04-R 012

音声の知的処理に関する調査研究
報　告　書

平成５年３月

JIPDEC

財団法人　日本情報処理開発協会

この報告書は、日本自転車振興会から競輪収益の一部である機械工業振興資金の補助を受けて平成４年度に実施した「音声の知的処理に関する調査研究」の成果を取りまとめたものである。

はじめに

音声をコンピュータとのインタフェースとして利用できれば、コンピュータを利用できるユーザの範囲は飛躍的に増大する。

最近、音声認識技術のうち音韻のモデル化の側面でいくつかの大きな進展があり、これまで非常に困難とされてきた、大語彙、不特定話者及び連続音声を考慮した音声認識システムの実現化が図られつつある。

音声処理については、連続音声認識技術の研究環境が整備されないため、銀行の残高照会、郵便物配送指示システム、クレーンの音声制御など、単語認識技術に基づく応用システムや、極めて語彙の少ない分野に絞った応用システムの開発に止っている。

実用レベルでの音声認識、自然な会話によるコンピュータとの対話を実現するための音声対話システムなどの技術開発の方向はまだ明らかでなく、そのための基盤（音声データベース等）さえ充分には整っていない。音声認識は、その対象とする情報が自然言語であり、欧米の技術をそのまま利用することはできないため、我が国独自の調査開発、実用化のための努力が必要になっている。

このような状況のもと、ＡＩ技術の振興を目的としているＩＣＯＴ－ＪＩＰＤＥＣ ＡＩセンターは、平成３年度から２年間の予定で、音声の知的処理に関する調査研究を計画した。本調査研究の実施にあたっては、知的音声処理調査研究委員会〔委員長：田中和世電子技術総合研究所知能情報部音声研究室室長〕を組織し、２年間に２０回の会合を重ねるとともに延べ６人の海外調査員を派遣するなど精力的に活動した。

本報告書は、平成４年度調査研究成果を取りまとめたもので、音声認識・理解技術を応用するためのシステム化技術、音声認識・理解技術の将来の応用分野、応用のための技術課題から構成され、当該分野の応用システム開発の指針となることを狙いとした。

なお、音声認識関連技術、応用システムおよび音声データベースの動向等については、平成３年度報告書に詳しくまとめたので併せてご参考頂ければ幸いである。

最後に、本調査研究の実施に当たり、御指導・御協力いただいた上記調査研究委員会の田中委員長をはじめ委員各位、関係各位に深甚なる謝意を表する次第である。

平成５年３月

財団法人　日本情報処理開発協会

財団法人 図書 日本情報処理開発協会

知的音声処理調査研究委員会・名簿（５０音順・敬称略）

委員長	田中　和世	電子技術総合研究所 知能情報部 音声研究室 室長
委　員	天野　明雄	㈱日立製作所 中央研究所 知識システム部 研究員
委　員	市川　　熹	千葉大学 工学部 情報工学科 教授
委　員	大洞　恭則	キヤノン㈱ 情報システム研究所 知能工学11研究室 主任研究員
委　員	木村　晋太	㈱富士通研究所 マルチメディア研究部 第３研究室 室長
委　員	鹿野　清宏	日本電信電話㈱ ヒューマンインタフェース研究所 主席研究員
委　員	徳永　健伸	東京工業大学 工学部 情報工学科 講師
委　員	長尾　　確	㈱ソニーコンピュータサイエンス研究所 アソシエートリサーチャー
委　員	中島　邦男	三菱電機㈱ 情報電子研究所 マルチメディア開発部 次長
委　員	中村　　哲	シャープ㈱ 技術本部 情報技術開発センター 第１開発部 主任
委　員	新田　恒雄	㈱東　芝 情報処理・機器技術研究所 開発第６部 課長
委　員	二矢田勝行	松下技研㈱ 研究開発グループ 主幹技師
委　員	速水　　悟	電子技術総合研究所 知能情報部 音声研究室 主任研究官
委　員	平山　　輝	大阪ガス㈱ 開発研究所 情報通信技術開発チーム 副課長

委　員	松本　　弘	信州大学 工学部 電気・電子工学科 助教授
委　員	渡辺　隆夫	日本電気㈱ C＆C情報研究所 メディアテクノロジィ研究部 研究課長
協力者	石川　　泰	三菱電機㈱ 情報電子研究所 マルチメディア開発部 主事
協力者	中川　聖一	豊橋技術科学大学 情報工学系 教授
協力者	畑崎香一郎	日本電気㈱ C＆C情報研究所 メディアテクノロジィ研究部 主任
協力者	星見　昌克	松下技研㈱ 研究開発グループ 主任技師
オブザーバ	佐伯　俊則	通商産業省 機械情報産業局 電子政策課 課長補佐
オブザーバ	西川　　勇	通商産業省 機械情報産業局 電子政策課 技術係長
オブザーバ	松村　　博	㈱メディアドライブ研究所 開発本部 部長
事務局	市川　　隆	(財)日本情報処理開発協会 ＡＩ・ファジィ振興センター 所長
事務局	片岡　幸一	(財)日本情報処理開発協会 ＡＩ・ファジィ振興センター 調査課 専任調査役

技術課題グループ・名簿（５０音順・敬称略）

主　査	田中　和世	電子技術総合研究所 知能情報部 音声研究室 室長
委　員	天野　明雄	㈱日立製作所 中央研究所 知識システム部 研究員
委　員	市川　　熹	千葉大学 工学部 情報工学科 教授
委　員	大洞　恭則	キヤノン㈱ 情報システム研究所 知能工学11研究室 主任研究員
委　員	木村　晋太	㈱富士通研究所 マルチメディア研究部 第３研究室 室長
委　員	鹿野　清宏	日本電信電話㈱ ヒューマンインタフェース研究所 主席研究員
委　員	徳永　健伸	東京工業大学 工学部 情報工学科 講師
委　員	長尾　　確	㈱ソニーコンピュータサイエンス研究所 アソシエートリサーチャー
委　員	中島　邦男	三菱電機㈱ 情報電子研究所 マルチメディア開発部 次長
委　員	新田　恒雄	㈱東　芝 情報処理・機器技術研究所 開発第６部 課長
委　員	松本　　弘	信州大学 工学部 電気・電子工学科 助教授
委　員	渡辺　隆夫	日本電気㈱ Ｃ＆Ｃ情報研究所 メディアテクノロジィ研究部 研究課長
事務局	市川　　隆	(財)日本情報処理開発協会 ＡＩ・ファジィ振興センター 所長
事務局	片岡　幸一	(財)日本情報処理開発協会 ＡＩ・ファジィ振興センター 調査課 専任調査役

応用グループ・名簿（５０音順・敬称略）

主　査	市川　熹	千葉大学 工学部 情報工学科 教授
委　員	田中　和世	電子技術総合研究所 知能情報部 音声研究室 室長
委　員	徳永　健伸	東京工業大学 工学部 情報工学科 講師
委　員	長尾　確	㈱ソニーコンピュータサイエンス研究所 アソシエートリサーチャー
委　員	中村　哲	シャープ㈱ 技術本部 情報技術開発センター 第１開発部 主任
委　員	二矢田勝行	松下技研㈱ 研究開発グループ 主幹技師
委　員	速水　悟	電子技術総合研究所 知能情報部 音声研究室 主任研究官
委　員	平山　輝	大阪ガス㈱ 開発研究所 情報通信技術開発チーム 副課長
オブザーバ	松村　博	㈱メディアドライブ研究所 開発本部 部長
事務局	市川　隆	(財)日本情報処理開発協会 ＡＩ・ファジィ振興センター 所長
事務局	片岡　幸一	(財)日本情報処理開発協会 ＡＩ・ファジィ振興センター 調査課 専任調査役

目　次

はじめに

第１部　調査研究の概要 ---------- 1
　第１章　調査研究の目的 ---------- 1
　第２章　調査研究項目および実施方法 ---------- 3
　　２.１調査研究項目 ---------- 3
　　２.２実施方法 ---------- 3
　第３章　調査研究の要約 ---------- 7
　　３.１音声認識･理解システム化技術 ---------- 7
　　３.２音声認識･理解技術の諸課題 ---------- 15
　　３.３海外動向 ---------- 22

第２部　音声認識･理解システム化技術 ---------- 27
　第１章　概　要 ---------- 27
　第２章　音声認識･理解応用システムの分類 ---------- 31
　　２.１はじめに ---------- 31
　　２.２音声認識･理解の機能 ---------- 31
　　２.３音声認識･理解による効果 ---------- 33
　　２.４音声認識･理解の形態と性能 ---------- 35
　　２.５音声認識･理解応用システムの特性 ---------- 38
　　２.６音声認識･理解応用システムの分類 ---------- 40
　　２.７音声認識･理解システムの現状 ---------- 43
　　２.８おわりに ---------- 47
　第３章　音声対話アプリケーション開発支援ツールの提案 ---------- 49
　　３.１ 音声対話アプリケーション開発支援ツールの必要性 ---------- 49
　　３.２ 音声対話アプリケーションの例 ---------- 49
　　３.３ 音声対話アプリケーション開発支援ツールの構成例 ---------- 52
　　３.４ タスク／ユーザモデルの構成例 ---------- 56
　第４章　現状音声認識技術適用時の留意事項 ---------- 63
　　４.１ 音声認識応用システム開発のガイドライン ---------- 63
　　４.２ 使用語彙の選択支援ツール ---------- 71
　第５章　音声認識･理解システムの評価 ---------- 75
　　５.１「ユーザの満足度」について ---------- 75
　　５.２まとめ ---------- 81
　第６章　音声処理技術の標準化 ---------- 83
　　６.１音声処理技術標準化の背景 ---------- 83
　　６.２マルチメディア標準化の動向 ---------- 84

6.3 音声処理技術標準化の将来 ---------- 86
第7章　今後の課題 ---------- 89

第3部　音声認識・理解技術の諸課題 ---------- 91
第1章　概　要 ---------- 91
1.1 はじめに ---------- 91
1.2 応用システムの概要 ---------- 92
1.3 応用システムからみた要素技術の研究課題 ---------- 93
第2章　要素技術 ---------- 99
2.1 人工知能要素技術 ---------- 99
2.2 ユーザインタフェース ---------- 117
2.3 話者認識 ---------- 134
2.4 音声規則合成 ---------- 153
第3章　応用システム別の技術課題 ---------- 165
3.1 音声ワープロ ---------- 165
3.2 福祉ロボット ---------- 177
3.3 CAIシステム ---------- 184
3.4 情報案内システム ---------- 191
3.5 予約システム ---------- 209
3.6 計算機用音声インタフェース ---------- 216
3.7 携帯端末 ---------- 225
3.8 マルチメディア（音声）データベース検索システム ---------- 231
3.9 ナビゲーションシステム ---------- 235
第4章　まとめ ---------- 237

第4部　音声認識・理解技術の海外動向 ---------- 239
第1章　米　国 ---------- 239
1.1 概　要 ---------- 239
1.2 訪問研究機関 ---------- 240
第2章　欧　州 ---------- 245
2.1 概　要 ---------- 245
2.2 訪問研究機関 ---------- 245
第3章　ICSLP 92 ---------- 251
3.1 会議の概要 ---------- 251
3.2 分野毎の研究発表概要 ---------- 253
3.3 音声データベースと評価の国際協調に関するワークショップ ---------- 254

資料編
A.「研究用連続音声データベース」利用ツール ---------- 257

第1章　概　要 ---------- 257
第2章　研究用連続音声データベース ---------- 259
2.1はじめに ---------- 259
2.2音声データベースの内容 ---------- 259
2.3CD-ROMフォーマット ---------- 260
2.4ディレクトリ構造 ---------- 261
第3章　ハードウェア構成およびツールの機能 ---------- 269
3.1ハードウェア構成 ---------- 269
3.2ツールの機能 ---------- 269
第4章　ツールの操作概要 ---------- 281
4.1はじめに ---------- 281
4.2機能概要 ---------- 281
4.3画　面 ---------- 283
4.4操作説明と仕様 ---------- 284
4.5ラベルファイルデータ形式 ---------- 297
4.6ラベルファイルデータ詳細説明 ---------- 298
B. 音声認識関連製品 ---------- 301
C. 音声合成関連製品 ---------- 309
D. 入手可能な音声データ ---------- 321

索　引 ---------- 323

第１部　調査研究の概要

第１章　調査研究の目的

第１章　調査研究の目的

音声は人間にとって最も基本的なコミュニケーションの手段であり、音声を用いてコンピュータと直接、対話することができるようになれば、マンーマシンインタフェースの飛躍的な向上が期待できる。コンピュータのパーソナル化、マルチメディア化などが今後大きく進展することが予想され、音声言語処理技術もこれらの一環としてその重要性がますます増大するものと見込まれる。

音声認識･理解技術は、最近の計算機性能の向上に伴って従来では困難であった計算規模、メモリ容量などが使用可能になり、その研究開発が活発化している。これによって、従来、特に困難な問題とされてきた大語彙、不特定話者音声、連続音声認識の分野へと急速に研究開発が展開されつつある。

そこで、本委員会では昨年度報告書において、このような音声認識･理解関連技術の動向を、

１）要素技術の研究開発動向

２）応用システムの開発ならびに利用動向

３）音声データベースの動向

に分けて調査報告した。

一方、音声認識･理解技術を応用の観点から見ると、その要求される技術水準はしばしばきわめて高度であることが多く、音声認識･理解技術の応用が将来に向けて必ずしも平坦であるとはいえない。したがって、その技術の普及と発展を促進するには、現実に音声認識･理解技術を応用するためのシステム化技術を明らかにしていく必要がある。また、より長期的には音声認識･理解技術がどのような分野に応用され、そのためにはどのような技術課題が存在するかなどを明らかにすることが重要である。本年度は、昨年度の報告書を踏まえた上で、上記のような観点から応用システム開発者への指針と将来の応用動向から見た技術課題を明らかにすることを目的とし、以下の３項目について、調査検討を行った。

（１）音声認識･理解システム化技術

（２）音声認識･理解技術の諸課題

（３）音声認識･理解技術の最新動向（海外調査）

第２章　調査研究項目および実施方法

第２章　調査研究項目および実施方法

２.１　調査研究項目

（１）　音声認識･理解システム化技術

音声認識･理解システムの分類をおこなうとともに、音声認識･理解技術を応用システムに組み込むための支援ツール、応用システムの評価方法、インタフェース条件やユーザインタフェースの使用手順の標準化等、応用システム開発における問題点に対応するためのいくつかの施策を検討した。

また、現在利用できる音声認識･理解技術を応用システムに組み込む際の注意すべきノウハウ的留意事項について整理をおこなった。

（２）　音声認識･理解技術の諸課題

昨年度の調査研究報告書では扱わなかった① 人工知能要素技術、② ユーザインタフェース、③ 話者認識、④ 音声規則合成　の４つの要素技術を取り上げて、その研究開発動向と技術課題を検討した。また、応用システム側からの技術ニーズを探るため、将来（おおむね10年後を目途に実現が見込まれる）応用システムを類型化し、９種類の応用システムを取り上げて、その利用形態や必要となる技術課題等について検討した。

（３）　音声認識･理解関連技術の最新動向（海外調査）

諸外国の音声認識･理解関連技術および応用システムの最新情報を収集するため、米国・カナダおよび欧州に調査員を派遣した。米国調査では、カーネギー・メロン大学、ハーバード大学、スタンフォード研究所などの訪問と国際会議 ICLSP'92（カナダ開催）などへ参加し、また、欧州調査では、ロジカケンブリッジ、ブリュストル大学、信号レーダー研究所（以上、英国）、国立情報処理研究所、フィリップスアーヘン研究所（以上、ドイツ）、国立情報制御研究所、国立情報機械研究所（以上、フランス）などを訪問し、当該分野の最新動向について調査した。

２.２　実施方法

上記の調査研究項目の実施にあたっては、当該分野の研究者、技術者より構成される「知的音声処理調査研究委員会」を当協会内に設置し、調査活動をおこなった。

平成４年度・委員会活動経過 (1/2)

月　日	検　討　テ　ー　マ	備　　考
平成４年 6月	6月24日　第１回委員会 1. 委員自己紹介 2. 平成４年度の活動計画について	
7月	7月22日　第２回委員会 1. 応用課題の検討方針について 2. 開発（音声DB利用）支援ツールについて 7月23日　第３回委員会 1. 技術課題の検討方針について	
9月	9月17日　第４回委員会 1. 応用システムの将来動向と技術課題の検討 2. 音声DB利用支援ツール（外部仕様）について 3. その他（米国・カナダ調査計画） 9月22日　第５回委員会 1. 応用システムの分類 2. 応用システム開発における課題の検討 3. 音声DB利用支援ツール（外部仕様）について	
10月	10月28日　第６回委員会 1. 米国・カナダ調査報告 2. 技術課題および応用課題の中間報告 4. 取りまとめ方針について 5. その他（欧州調査計画）	10月10日～ 10月24日 米国・カナダ調査 （田中委員長）
12月	12月10日　第７回委員会 1. 技術課題編（報告書）の構成と内容について 12月16日　第８回委員会 1. マルチメディアに関する標準化の動向 2. 応用課題編（報告書）の構成と内容について 3. 欧州調査報告 4. 音声DB利用ツール開発状況の報告	11月28日～ 12月11日 欧州調査 （速水委員）

平成４年度・委員会活動経過 (2/2)

月　日	検　討　テ　ー　マ	備　　考
平成５年 2月	2月2日　第９回委員会 1. 応用課題編（報告書）原稿案の検討 2月4日　第10回委員会 1. 技術課題編（報告書）原稿案の検討	
3月	3月3日　第11回委員会 1. 報告書（案）の検討	

第３章　調査研究の要約

第３章　調査研究の要約

３．１　音声認識･理解システム化技術

３．１．１　音声認識･理解応用システムの分類

音声認識･理解を実際に応用システムに適用するためには、音声認識･理解技術を導入するメリットやデメリット、導入コストや現在の技術レベルを知ることが不可欠である。音声認識･理解技術は、現在のところ万能でなくコストも高く、現在の技術レベルの範囲で実現可能か、どのように構成するかをよく検討し導入に移らなければならない。

第２部２章では、音声認識･理解の機能、効果、システム特性、制約、現状技術について音声認識･理解装置実用化の成功例を見ながら、音声認識･理解応用システムの開発、導入の目安となるよう音声認識･理解応用システムの分類とその特性の整理をおこなった。

（１）音声認識･理解の機能

音声認識･理解の機能には、種々の機能があるが応用に関していえば、音声認識･理解でないといけない、音声認識･理解だからこそ実現する機能というものに注目する必要がある。容易に代替手段が存在する機能を無理に音声で実現するのは合理的ではない。このような音声認識･理解独自の特徴的機能としては、非接触制御機能、電話回線入力機能、多項目入力機能、音声対話機能の大きく分けて４つのものが挙げられる。

（２）音声認識･理解による効果

音声入力の機能を生かしてシステムを構成することにより、システムとしての効果が生まれる。これは、実際に使用者に与えるメリットというよりむしろ導入者に対するメリットともいえる。音声認識･理解応用システムを開発するためには、開発したシステムを導入してもらう必要があり、導入者のメリットを考えることが重要となる。ここでは、音声認識･理解による効果として、即時化、省力化、ユーザフレンドリー化、高信頼性化、低価格化、宣伝効果の６つの効果が挙げられる。ただし、これらの６つの効果は、独立ではなく多かれ少なかれ相関がある。

（３）音声認識･理解の形態と性能

システムを構築するために必要となる音声認識･理解技術の条件、音声認識･理解の形態と性能について述べる。

A.　発声単位

使用者が認識装置に発声する発声の単位を指している。発声の単位には、大きく分けて次

の6種類があり、下にいくほど難易度が高い。機能としては連続単語まで実用化されているが、連続単語モードでは性能が十分でなくあまり使われてないという調査結果もある。

・孤立単語発声
・キーワード発声
・連続単語発声
・文節発声
・読み上げ発声
・自由発話

B. 発声形式

Aで述べた発声単位をどのように発声するかの問題である。装置によっては、色々な発声がなされ難易度は変わる。このため、発声のスタイルは1つの条件とみなすことができる。次の3つのタイプを考慮する必要があり、下にいくほど難易度が高い。また、雑発声に対しては現在の認識装置では性能が大きく劣化するのが実状である。

・丁寧発声
・自然発声
・雑発声

C. 話 者

どのような話者を対象にするか。大きく分けて、次の4つがあり、下にいくほど難しい。話者といっても声の音色の違いに加えて、話し方、方言の違いまであるので実際には不特定話者を認識するのは極めて難しい。現状では、話者に関しては認識語数との関係が強いので一概にどのレベルまで可能かを述べることは困難であるが、商品レベルでは100単語以上なら特定話者、それ以下なら不特定でも可能というところである。

・特定話者
・複数特定話者
・話者適応 (使用前に少し発声してもらえ、ある程度の間使用する)
・不特定話者 (いきなり発声する)

D. 語 彙

どのような語彙を設定するかということである。あらかじめ決まった語彙なのか、使用者が

設定できるかなど分けると次の3つのタイプがある。下にいくほど難易度が高い。

・特定語彙
・可変語彙
・任意語彙

E. 語　数

認識語数は、応用に応じて設定される。不特定話者に対応する場合は、100単語以下が適当であろうし、特定話者なら100単語以上可能だが登録が大変になる。このためシステムの制約と、認識性能との妥協点で設計することになる。

F. マイクロホン

使用する環境、使用形態、システムの構成によりマイクロホンの形態が決定される。音声を発声する口とマイクロホンの距離が近いほど認識性能を高くすることができるが、この距離の近い接話マイクロホンは話者を拘束するので使用者に好まれない。また、システムによっては、ワイヤレスの接話マイクロホンを用いるものもある。

G. 騒音環境

騒音環境は、それ自体音声の特徴パターンに歪みを与え認識性能を劣化させるばかりでなく、環境により音声パターンに発声変形を与える。これにより、認識性能は大きく劣化する。このためどのような環境において使用されるかで、認識方法、使用するマイクロホン、システム全体の設計を検討する必要がある。 大まかに分けて次のような騒音環境があり、下にいくほど難しいと考えれる。現在音声認識技術開発において対象とされているのは、家庭内、自動車内、駅の改札口、オフィスレベルなどである。

・家庭内環境
・オフィス環境
・自動車内(一般道路、高速道路、窓開け、オープンカー、雪道、トンネル)
・工場
・屋外(郊外、近郊、道路沿い、パチンコ店、駅のホーム…)

（4）音声認識･理解応用システムの特性

音声認識･理解を応用してシステムを構成するにあたって考慮すべきこととして、インタフェースの関係、情報の向き、主導権、タスクの限定可能性がある。これによってシステム

の構築の難易度が変わる。

表３．１-１に、音声認識･理解応用システムの分類とその特性について示す。

表３．１-１　音声認識･理解応用システムの特性

分　類	概要・機能	特　徴	応用例
A.　同定型	個人、 言語の同定	信号レベルの同定、 機械主導、 非接触、 即時	話者識別キー
B.　項目入力型	多項目入力	単語認識、 機械主導（人主導）、 特定話者、 非接触、 電話回線、 即時	荷受データ入力、 検査システム、 肉等級入力システム、 配送仕分システム、 歯科検査記録システム
C.　命令型	音声制御、 音声リモコン	単語認識、 機械主導（人主導）、 不特定話者、 非接触、 電話回線、 即時	音声認識電話、 音声認識券売機、 トーククレーン、 ビデオ予約リモコン、 ANSERシステム
D.　教育型	人間に知識を供与する （CAI）	対話理解、 話し言葉、 機械主導、 音声対話、 機械が人間を教育	英語CAIシステム
E.　解釈型	音声認識合成　＋ エキスパートシステム	対話理解、 話し言葉、 人主導	
E-1.　相談型	人間が理解するための 問い合わせ	音声対話、 人主導、 機械から提案	電話番号案内
E-2.　解釈、計画型	機械が理解しまとめる （申込み、予約）	音声対話、 機械主導、 機械がまとめる	旅行プラン作成システム
E-3.　変換、翻訳型	音声言語の翻訳、 認識符号化	対話理解、 人主導、 翻訳	翻訳電話
F.　ディクテーション型	文書作成のための入力	文章認識、 書き言葉、 人主導、 非接触、 高速入力、 タスクフリー	音声ワープロ、 ディクテーション システム

3.1.2　音声対話アプリケーション開発支援ツールの必要性

音声認識·理解システムの利用拡大を考える上で、認識率や耐騒音性など音声認識技術の性能向上に加えて、一般のシステム開発者やユーザが現状の音声認識技術を用いて、実用的なアプリケーションを簡単に開発できるような環境を提供することが重要であると考えられる。特に、音声アプリケーションには多かれ少なかれ必ず存在するシステムとユーザ間の対話制御に関しては、簡単なツールを提供しない限り、数多くの音声アプリケーションがユーザの手によって開発されていく状況を期待することは難しい。もちろん、自然で自由な発話による音声対話など高度で複雑なシステムの実現は非常に難しく、今後のこの分野の研究の進展を待たねばならない。しかしながら、アプリケーション分野のなかには、前項で述べた音声認識·理解応用システムの分類の内、項目入力型、命令型、教育型、解釈・計画型など、機械主導型の非常に限定された会話でその目的が達成でき、その対話の流れも状態遷移で記述できるような比較的簡単なものが数多くある。そこで、このようなアプリケーション用に、音声認識や言語処理の研究開発者ではなく、アプリケーション開発者やユーザが、対話の制御を比較的簡単に定義できるツール、「音声対話アプリケーションビルダー」を提案する。

音声対話アプリケーションの開発に一般的な手順が確立されているわけではないが、連続音声認識による対話システムを想定すると、そのアプリケーション開発には次のようなステップが考えられる。

（1）アプリケーションの選定
（2）対話例の採取
（3）対話の流れの定義
（4）語彙、文法の定義
（5）システム全体の開発、他のインタフェースとの統合
（6）テストと評価

音声対話アプリケーションビルダーは、これらすべてのステップにおいて、アプリケーション開発者を支援するものであることが望ましい。本調査研究では（3）、（4）、（5）を支援するものを第2部3章で提案し、（6）については、第2部第4章で述べた。

3.1.3　音声認識応用システム開発ガイドライン

現在利用可能な音声認識·理解技術を、実際に応用（アプリケーション）システムに組み込む際に注意すべきノウハウ的留意事項について整理をおこない、第2部4章にまとめた。

音声認識応用システムの開発においては、3.1-2に示したようなさまざまな項目について

考慮し、その応用に適した仕様を検討する必要がある。

表3.1-2　音声認識応用システム開発における留意事項

---

(1) 音声認識の形態 1)~3)

1) 話者：特定話者／話者適応／不特定話者

2) 発声単位：孤立単語／連続単語／文節／文

3) 語彙：認識対象語数、単語の選択と組合せ

---

(2) タスクの設計における留意事項 4)~6)

4) 文法、文型の定義

5) 対話制御の設計

6) ユーザインタフェース

---

(3) その他の留意事項 7)~9)

7) 収録系

8) システム評価法（テスト法）

9) ハード構成 (音声認識ボード／音声認識装置／音声認識 LSI)

3.1.4　音声認識·理解システムの評価

音声認識·理解システムを評価するための評価項目について提案を行う。ユーザは実際に使用するエンドユーザと、システムを導入するユーザの2つがある。エンドユーザは”使いやすさ”で満足度を判断するが、システムの導入者は音声認識·理解システムを導入することによって”仕事が経営的にどの程度改善されるか”で満足度を判断する。

表3.1-3にエンドユーザの意見をまとめる場合の評価項目、表3.1-4にシステム導入者に対する評価項目を示す。各項目に関してどの程度重視するかは場合によって異なると考えられるので、ここでは触れないことにする。

表3.1-3　ユーザに対する満足度評価項目

(1) 目標達成に関して

(a) 目標達成までの時間	平均	秒
(b) 目標達成までのｲﾝﾀﾗｸｼｮﾝ数	平均	回
(c) 目標達成率	平均	%
(d) 目標に対してどの程度情報が得られたか	平均	%

(2) 発声に関して

(a) 発声のタイミング	自然			不自然
(b) 不要語、言い淀みの発声	可			不可
(c) 発声の言葉に対する制限	無			有

(3) 認識性能

(a) 認識性能	満足			不満足
(b) 言葉による認識率のばらつき	少ない			多い
(c) 話者による認識率のばらつき（最高と最低認識率の差）				%

(4) 使いやすさ

(a) 使いやすさ	良			悪
(b) マニュアル（説明）なしで使用できる	可			不可
(c) 状態遷移 (Q&Aなど) の自然さ	自然			不自然
(d) 常にどの状態にいるかがわかりやすい	容易			難
(e) 誤認識した場合の修正	容易			難

(5) 感想

(a) また、使いたいと思う人の割合　　%

(b) 感想

表3.1-4　システム導入者に関する項目（経営的に）

(6) システム導入者に関する項目（経営的に）

(a) 人の削減（人件費の削減）	大				小
(b) 仕事のコストパフォーマンスの向上	大				小
(c) 仕事の信頼性の向上	大				小
(d) 会社の宣伝効果（イメージアップ）	大				小

(7) 感想

3.1.5　音声処理技術の標準化

コンピュータの世界は、長い間、数値・記号データを処理の対象としてきた。しかし、処理速度・メモリ容量が飛躍的に向上したのに伴い、近年、処理対象を他のメディアに拡大する試みが進んでいる。人間同士の対話が様々なメディアを使用していることを考えると、コンピュータとの対話チャンネルがマルチメディア化することは、自然な流れである。他方、デジタル通信の世界においてもネットワークの高速・大容量化に伴って、音声・文字・画像のマルチメディアを同時に伝送する、デジタルネットワーク統合サービス（ISDN）の整備が進められている。ここでは処理の対象が音声から、文字・画像を含むマルチメディアに拡大されるに伴い、ATM（非同期伝送モード）などの新しい技術が開発されている　処理の対象がマルチメディアへと拡大された結果、データフォーマットやプロトコルに関する規約を決める必要が生じている。そこで、第2部6章では、マルチメディア処理技術や音声処理技術の標準化の動向と、その将来について解説する。

3.2　音声認識·理解技術の諸課題

3.2.1　将来の応用システムと技術課題

音声認識·理解技術の研究開発動向、技術課題については、昨年度報告書において詳しく論述してきた。しかし、その視点は技術開発側からのもであり、応用システム側からの技術ニーズの視点に立ったものではなかった。そこで、今年度は、将来の音声認識·理解技術の応用動向を予測して、応用システム側からの視点で必要となる要素技術、研究開発課題について考察した。

（1）要素技術

音声認識·理解応用システムに必要な要素技術のうち、昨年度報告書に含まれていない以下に挙げた4つの技術の研究開発動向について考察した。

1）人工知能要素技術

2）ユーザインタフェース

3）話者認識

4）音声規則合成

人工知能要素技術は、将来の高度な音声認識·理解システムにとってきわめて重要であるばかりでなく、音声認識·理解技術の応用範囲を拡大する上でも重要な役割を担うものと期待される。ユーザインタフェースは、利用者のユーザビリティを向上させる重要な技術であるが、これまで比較的研究が少なかった分野である。今後、拡大が見込まれるマルチメディア、マルチモーダル・インタフェースなどと併せて活発な研究開発が望まれる。話者認識技術は、多くの実際的な応用局面において利用者（話者）を限定したいという要求があり、必要となる技術である。音声規則合成は、音声応用システムにおいて重要な位置を占める音声応答機能のための基本技術である。

（2）応用システム別の技術課題

おおむね10年後を目途に将来実現が見込まれる応用システムを類型化し、以下の9種類の応用システムについて想定される利用形態や、必要となる技術課題について考察した。

1）音声ワープロ　　2）福祉ロボット

3）ＣＡＩシステム　　4）情報案内システム

5）予約システム　　6）計算機用音声インタフェース

7）携帯端末　　8）マルチメディアデータベース検索システム

9）ナビゲーションシステム

これらの応用システムは、音声認識·理解システムの応用を機能要素に分類し、バランスよく選択したものである。

音声認識·理解技術の応用を想定したとき考えられる基本的な機能として以下のようなものが挙げられる。

a）指示、命令（たとえば、ロボットに対する音声による指示など）
b）案内、問い合わせ（各種情報についての問い合わせ、データベースからの情報サービスなどに音声入力を利用）
c）予約（電話などの音声による各種の予約、注文など。案内との違いは、システム側のデータベースの書換えを伴う点）
d）文書作成（音声による定型文書や議事録などの作成）
e）教育、CAI（知的CAIへの音声入出力の組み込み、管制訓練、語学練習など）
f）自動通訳（音声理解、自動翻訳、音声合成を組み合わせたシステム）
g）汎用インタフェース（計算機に組み込まれた汎用の音声入出力）

以下に、応用システム別の利用形態と研究開発課題の概要を述べる。

1）音声ワープロ

音声を主入力機能とするワープロで、特定業務文書作成、自由文書作成、速記メモなどへの利用が考えられる。ユーザフレンドリなインタフェース、大語彙認識などの機能が望まれる。

2）福祉ロボット

高齢化社会の到来などを考慮すると、看護、介助等にロボットを利用する必要性が高まり、同時にこれらを音声により制御するという機能要請も増大すると考えられる。高齢者や身体の不自由な人などの多様な音声への適応が課題となる。

3）知的CAI（音声インタフェース機能）

知的CAIにとって音声言語インタフェースは必須である。利用者（学習者）は、システム側からの質問に対し思考による沈黙や質問、躊躇することなどが考えられるので、言い淀み、言い直しなどに対する言語処理が望まれる。

4）情報案内システム（マルチメディア）

計算機の高性能化によって、大量の情報を整理してデータベースとして蓄積することが可能になりつつある。この情報ベースから、音声を含むマルチモーダル入力・マルチメディア出力によって、ユーザが有益な情報提供を受けるシステムが考えられる。

5）予約システム

現在のところ、音声認識機能自体を予約に用いているシステムは見あたらないが、将来的には、発達した通信ネットワークを利用して、携帯端末や移動端末等からの音声による各種予約、買物注文などが考えられる。普及のためにはインフラストラクチャの整備、バーチャルリアリティ技術なども課題となる。

6）計算機用音声インタフェース

個人向けのパーソナルコンピュータやワークステーションの普及とそれらのマルチメディア機能化に伴って、音声インタフェースを組み込むことが考えられる。ここでは、アプリケーションプログラムとのインタフェースの標準化が重要な問題となる。

7）携帯端末

装置のハード面から考えると、音声入力を主体とするキーボードレスに近い仕様によって、携帯性のよい端末が考えられる。利用範囲は各種考えられるが、インフラ整備が重要となる。

8）マルチメディアデータベース（DB）検索システム

今後、普及増大すると見込まれるマルチメディアDBを音声で検索するシステムで、ユーザインタフェースに優れる利点がある。特に、音声DBファイルを直接音声で検索するシステムは、DB構築やDB管理システムの負担軽減につながる。

9）ナビゲーションシステム

自動車への音声認識関連技術の応用は、運転中、手が離せないという状況を考慮すると有望な分野の1つと考えられる。ここではその代表として音声インタフェース機能をもつナビゲーションシステムを取り上げる。

3.2.2　応用システムからみた要素技術の研究課題

昨年度報告書でその研究開発動向を詳述した音声認識·理解要素技術について、改めて簡単に復習し、これらの技術が将来の応用システムのどのような局面において重要となるかという視点から考察する。要素技術には、大語彙化のための認識技術、連続音声や自由発話、対話音声を扱う技術、話者適応化や不特定話者対応の技術、話者識別、発話環境（耐雑音、電話音声、低品質音声）処理技術などがある。また、音声認識·理解システム関連技術として、音声合成技術やシステムのユーザインタフェース技術も重要である。

（1）大語彙音声認識

大語彙化は、一般的にはシステムの高度化や用途の拡大にとって欠かせない重要な要素である。また、システム構築において語彙の選択あるいは限定を無理なくおこなえること、利用者が語彙に関する制約を比較的意識せず利用できることなどの利点もある。しかし、一方では、

認識率や応答時間などとは相反する関係にあり、応用上はシステムの目的によく適合した語彙量を選ぶ必要がある。

大語彙化のための技術課題としては、主に、音声特徴抽出技術、有効な音韻モデルなどの音響音声学的レベルでの高性能化が挙げられる。特徴抽出技術については、最近ではほとんどシステムが音声のフーリエ・パワースペクトル包絡を用いていて、新しい傾向が見られない。大語彙化に向けてはこの部分でのブレークスルーが望まれる。音韻モデルに関しては、隠れマルコフモデル（HMM）によるモデル化が多く、その精密化が認識率の改善に寄与することが明らかになっているものの、モデルのロバストネスとのトレードオフが問題となっている。また、大語彙化に伴う処理時間の増大を回避するための高速検索技術、有効な辞書構成と実装技術などが必要となる。

大語彙は多くのシステムにとって望まれる機能であるが、用途を限定したシステムでは、むしろ認識率や未登録語のリジェクト、応答時間が重視される場合もある。大語彙が基本的に必要な応用システムとしては、一般的な文書作成のための音声ワープロ、案内システムなどでは数字、名前など語彙数が多くなる対象を扱うシステム、あるいはDB検索で項目数が多いシステムなどである。逆に、福祉ロボットやナビゲーションのようなシステムでは、大語彙化よりも即応性のような応答時間、マクロ命令への対応などが望まれる。

（2）連続音声認識、自由発話音声、対話音声処理

発話の様式については、認識率の面からは孤立単語発声が断然有利であるが、ユーザインタフェースを考えると、連続音声さらには自由発話音声への対応が望まれる。このためには、従来の自然言語処理技術に加えて曖昧な入力を処理できる音声言語向きの言語処理技術が必要となる。また、自由発話音声処理には、音声に通常含まれる非言語音、笑い、言い直しなどの非文法的発話、未知語（辞書にあらかじめ登録されてない語彙）などの処理技術が要求される。さらに、音声言語にはしばしば、意味的誤り、多重性なども含まれる。これらの問題の処理は、自然言語処理の分野でも最も困難とされてきたもので、音声との結合ではさらに複雑となる。ただし、これらの言語現象については、自然言語処理の面のみでなく、韻律など音声的側面からの研究も今後進めていくべき研究課題である。

自由発話音声や対話音声については、人工知能技術、特に対話モデルなどとの関連も強く、第３部２.１節における技術動向解説を参照されたい。これら課題についての音声研究の歴史は浅く、蓄積されている知識も少ない。有効な音声データベースの構築が当面の課題の１つである。最近の傾向として、統計量にもとづいた確率言語モデルなどの提案もあるが、そのために必要となるデータ量は莫大であり、現実的な解決をみたものは見あたらない。このためには、音声言語データの計量化法、自己組織化や構造化法、タスクの複雑さの評価法、音声理解シス

テムの性能評価法等についての研究も重要である。

上述したように、自由発話を許容することはユーザインタフェースの大幅な向上をもたらす。特に、案内やコンサルティング、予約システム、知的CAI等への応用についてはその実現が強く望まれる。しかし、認識性能、システム構築上の難度を考慮すると、自由発話の許容は不利であり、システムの確実、安定な動作という点では不安がある。一方、孤立単語発声や、文節単位発声でも使用上の著しい不都合はないシステムも考えられる。たとえば、音声ワープロ、単純な案内、指示・命令を主とするロボットのようなシステムである。

対話型のシステムは単純なものから高度なものまで考えられるので、対話音声の範囲はかなり広くなる。ロボットのような指示命令機能の場合には、単純な対話でも有用な場合が多い。これに対して、知的CAIのようなシステムでは、利用者の反応によってどのような応答を返すか等の指導方略に関連してきわめて高度な音声対話処理が考えられる。

（3）話者適応、不特定話者音声認識

音声認識装置が任意の話者の発話を何の訓練もなしに認識できれば、個々の話者の声を登録する方式に比べ、明らかに利便性に優れる。しかし、音声の音響物理的特性は同じ音声言語でも話者によって多様に変化するため、いわゆる不特定話者音声認識はきわめて困難な技術の1つである。次善の方式として、少ない音声サンプルを用いて、発話者の特性にシステムの適応化を図る話者適応化技術があり、最近活発に研究されている。話者適応化処理は、システムの性能を特定話者方式の性能とほぼ同等に引き上げることができ、応用面から現実的な方策の1つである。しかし、話者がきわめて多数でかつ適応化処理のような余裕が認められない類の応用も数多く考えられる。このような場合には、必然的に不特定話者音声を対象とするシステムが必要になる。

話者適応化技術の多くは、標準話者と入力話者とのベクトル符号の対応付け、特徴空間分布のパラメータの対応付けにもとづいている。いづれも静的ないし局所的な音韻的特徴の適応であり、より動的あるいは発話形態的特徴の適応などは今後の課題である。

不特定話者化の技術は、音韻表現モデルに若干の違いはあっても基本的には複数標準パターン化である。これは方式上、現実的な選択であるといえるが、任意の年齢や地域など広範な話者を考慮した場合、システムのロバストネスの点で依然として大きな問題がある。

話者適応化技術は、いずれ将来的には、いわゆる特定話者方式を包含するものになると考えられる。したがって、主に個人向け仕様のシステムや、利用者が少数に限定されるようなシステムに組み込まれるであろう。たとえば、音声ワープロや端末、ロボットなど。しかし、多くの不特定利用者の随時使用を想定するシステムでは、不特定話者音声認識が必要とされる。いわゆる学習によって短時間のうちに話者に意識させることなく適応化するというシステムは方

式としては考えられるが、応用を想定すると実用性が低いように思われる。

（4）話者認識

音声分析にもとづく話者認識には、個人（話者）の同定（照合）と話者識別があり、おおむね、指紋などと同じようにセキュリティへの応用の観点から研究されてきた。この機能は、音声を利用するというユーザフレンドリ性に優れ、多くの分野で応用可能性をもっているので研究はかなり長期にわたりおこなわれてきた。音声認識・理解装置への組み込みに関しては、装置が特定の利用者にのみ使用可能であるようにするという要請からきている。特に、装置側がリアルタイムに特定の話者のみの声に選別的に反応するならば、その効果はきわめて大きいといえる。

話者認識技術の現状と課題については、第3部2．3節で解説されるので、参照されたい。話者認識技術には、あらかじめ話者の話す言葉を知っている場合（テキスト依存）と、知らない場合（テキスト独立がある）。上記で述べたような使用状況を想定した場合は後者が要求されることが多く、しかもリアルタイム性を要求されると、現状の技術ではきわめて困難といえる。また、同じ個人であっても一般に時期（時間）や、健康状態、周囲環境による人間の声の不安定さは、よく知られるところである。機能としての利点は明かであっても、実現についての技術的困難も大きい。

（5）発話環境処理

各種の雑音、騒音環境など音声入力における発話環境処理は、実使用を想定するときわめて重要な要素技術である。騒音の種類は使用環境に依存して変わるが、使用環境を決定する要因として次のようなものが考えられる。まず装置の設置環境である。これには電話系や、携帯端末を利用するというものも含まれる。従来は、工場、作業場などでの使用が多かったが、今後の応用を見込むと家庭内、公共スペース、オフィス、自動車内などの環境での使用が増大すると考えられる。

雑音処理技術は、大きく分けて、マイクロホンなどの集音装置に工夫をする方式とソフト的に信号処理技術により雑音低減をおこなう方式がある。前者にはたとえばノイズキャンセリングマイクやマルチマイクの使用がある。後者としても様々な手法が提案されているが、雑音の性質が変化し易いものや非定常な雑音に対しては有効な手法の開発はきわめて困難である。

使用環境には、ハード的な側面とソフト的な側面からの問題がある。ハード面とはマイクと発話者との相対的な位置関係などで、マイクの種類もこれによってある程度限定されてくる。ヘッドセットのマイクは耐雑音性では有利であるが、一般には敬遠されることも多く、公共的な使用には適さない。また、ロボットのような場合、移動するロボット自身にマイクがセット

されるとすれば、処理系はきわめて困難な問題となる。ソフト面では、いわゆるロンバート効果のように、使用者（話者）が使用環境に影響されて、発話の変動が大きくなるという問題がある。この問題は、環境が特殊になればより大きく出てくる。たとえば、ささやき音声なども使用環境によっては想定する必要がある。

外部騒音や回線雑音のような雑音音声とは異なるが、高齢者や傷病者、障害を持つ人の音声処理も将来の応用を想定すると重要な課題である。

また、装置自体の耐性などの面での耐環境性、装置の仕様とその他設備、法規、社会環境との整合性などももちろん実際上は問題となる。

（6）音声合成

音声認識·理解機能を組み込んだシステムは、多くの場合、音声応答機能を有している。システム側からの応答は、一般には入力自体についての確認と入力に対する回答や動作に分かれるが、そのメディアとしては用途に応じて様々なものが考えられる。音声応答は音声が応答として便利ないし効率的な場合についての使用はもちろんのこと、システムの親しみ易さの増大にも貢献する。これらの音声応答は音声規則合成技術によって実現されている。音声規則合成は、入力として日本語文章テキストを想定した「テキスト-音声変換」の機能とも言えるが、応答の内容を表す概念を出発点と考えてもよく、この場合は「概念からの音声合成」と呼ばれる。

音声規則合成技術における問題点は、1）テキスト解析などの言語処理技術が不十分であること、2）韻律（アクセントやイントネーションなど）の自然性の確保、3）音質の自然性、肉声感の確保、などにある。最近の規則合成方式には、大量の音声素片を蓄積しておきその接続にもとづく方式が多く、音質の自然性などの面でかなり進歩してきている。しかし、音質自身の制御、韻律の制御など本来パラメータによって制御すべき要素については問題解決に至った訳ではない。音声の技術動向については第3部2．4節で詳述する。

音声合成は、応用面からみた総合的評価法の確立が重要であり、最近ようやくその観点からの研究が開始されつつある。すなわち、基本的な評価要素である明瞭性や自然性についても、場面や状況に応じてどのような音声が有効であるのかといった観点からの評価が重要である。したがって、単一（同質の）の音声を合成する手法ではなく、パラメータによって声質の制御ができる方式が将来望まれる方向である。

（7）ユーザインタフェース

システムの総合的評価はユーザインタフェースの良否に大きく依存するので、ユーザインタフェースの重要性は論を待たないが、比較的定量的な調査研究結果を出しにくい分野である。

このため、既存の技術や知見を踏まえて次の技術が登場するという側面が相対的に少ない。ユーザインタフェースを構成する技術要素には、表層的な操作性に依存する部分と、より深いたとえば人工知能技術や音声処理技術の本質的な進歩に依存する部分があり、それらがある時点で統合化されて1つのシステムとなる。

ユーザインタフェースの1つの切口としてモーダリティがある。将来的には明らかにマルチメディアを用いた情報提供とマルチモーダルな入力が優勢になると予想される。しかし、このようなマルチモーダルという点の有効性・有用性を明解に示した実験は必ずしも多くなく、今後の研究に負うところが大きい。これらの技術動向については第3部2.2節で詳述される。

ユーザインタフェースの向上のためには、ユーザのモデル化が重要な課題となる。将来的には、ユーザの応答から音響音声的情報を利用してユーザの感情などを推測するような情緒情報処理なども考えられる。また、ユーザインタフェースの評価法も重要な課題の1つである。

3.3 海外動向

3.3.1 米 国

米国を中心とする海外の音声認識・理解技術およびその関連技術の動向を調査するため、平成4年10月10日から2週間の日程で、カナダ、米国を訪問した。前半の週は音声言語処理に関する国際会議 International Conference on Spoken Language Processing（ICSLP）-92に参加し、世界の第一線の研究者の発表を聴講し、研究開発動向を調査した。後半の週は、昨年度の海外調査で訪問できなかった米国の先進的な研究機関であるCMU（Carnegie Mellon University）、ハーバード大学（Harvard University）、SRI（Stanford Research Institute）Internationalを訪問し、見学・意見交換をおこなった。本項では、米国の全体的な研究開発動向について述べ、3.3.3項でICSLP-92について報告する。

米国では、DARPAの下で音声言語理解プロジェクト、SLS（Spoken Language System）が強力に進められており、その予算規模は年間約1000万ドル余りといわれている。このプロジェクトにはMIT（マサチューセッツ工科大）、CMU（カーネギーメロン大）、SRI（スタンフォード研究所）、NIST（国立標準研究所）、BBN（ボルトベラクニューマン社）、TI（テキサスインスツルメント社）、AT&Tベル研究所などの音声研究グループが参加している。このプロジェクトでは、共通の音声データベースを構築し、さらに共通のタスクを設定して、各研究グループが競争的に研究をおこなっている。また、年1回のワークショップを開き研究発表と意見交換の場を設けている。このワークショップの論文集は、Proceedings of Speech and Natural Language Workshop（Morgan Kaufman Publishers）として市販されている。

3.3.2 欧 州

平成4年11月28日から12月11日にかけて、欧州における音声認識、音声理解および関連技術に関する動向調査を実施した。イギリスでは、ロジカケンブリッジ (Logica Cambridge)、ブリストル大学 (University of Bristol)、RSRE (Royal Signals and Rader Establishment) の3機関、ドイツでは、GMD/IPSI (Gesellshaft fur Mathematic und Datenverarbeitung / Integrated Publication and Information Systems Institute)、フィリップスアーヘン研究所の2機関、フランスでは、INRIA (Institute National de Recherche en Informatique et en Automatique)、LIMSI (Laboratoire d'Informatiquepour la Mechanique et les Sciences de l'Ingenieur) の2機関を訪問し、研究討論および情報交換をおこなった。これらの機関は、欧州における人工知能研究の代表的研究機関であり、特に音声認識、音声理解に関する研究分野では多くの先端的な研究がなされている。

研究分野の動向として、音声認識システムの性能が着実に向上していることが感じられた。各研究機関の音声認識システムの認識性能を比較するためのベンチマークテストが、米国DARPAの援助により開催されている音声と自然言語のワークショップでおこなわれてきたが、今回はフィリップスアーヘン研究所とLIMSIの2機関もこれに参加して高い認識性能を達成したとのことであった。

また音声対話を工学的に実現するための対話処理についての研究とともに、言語学的、認知心理学的な側面も含めた、音声対話についての幅広い研究が見られるようになってきている。

音声認識、音声理解技術の応用分野としては、ホームバンキング、列車旅行案内、自動車用ルート案内、オフィスレターの口述等のさまざまな分野で、音声認識、音声理解技術の適用が試みられていた。

欧州での音声認識、音声理解の研究は ESPRIT 計画のもとでもおこなわれてきたが、最近は予算的には伸び悩んでおり、全体的には研究費の確保がなかなか厳しい状況にあるようであった。

3.3.3 ICSLP '92

この国際会議は、8学会 (米国音響学会、日本音響学会、カナダ音響学会、カナダ言語学会、ヨーロッパ音声通信学会、IEEE信号処理学会、国際音声学会、米国言語学会) が提携して、カナダ・アルバータ大学の主宰で、アルバータ州バンフのBanff Springs Hotelにおいて、1992年10月12日～16日の期間開催された。この分野の国際会議としては世界最大規模である。今回が第2回目であり、第1回は1990年に神戸で開催され、それ以後、隔年ごとに開催されることで合意された。次回は1994年横浜で開催されることが決っている。

今回の会議では、合計442件の研究発表があり、このうち267件が口頭発表、残りがポスタ

ーでの発表であった。また、4件のplenary lectureがあった。全体のプログラムは表3.3-1に示す通りで、口頭発表は4つのセッションが並行しておこなわれ、さらにポスターセッションが1つ設けられているという形式であった。

会議参加者の人数はおおむね600人程度と見積られ、全体を通じて盛況であった。研究発表件数を国別に見ると表3.3-2のようになり、合計で27か国からの発表があったことになる。したがって、これらの国を中心に第一線の研究者が参加したものと見込まれる。参加者の中には、音声研究の分野で1950年代より活躍しているGunner FantやK. N. Stevensなど著名な研究者多数が含まれ、彼等自身が研究発表をおこなっていた。

表3.3-1　ICSLP 92　PROGRAM SCHEDULE

				1	2	3	4	Posters
TUESDAY	Morning	first	0830-1010	OPENING & PLENARY				
		second	1030-1210	Word Spotting in ASR	Speech Coding	Speech Production: Coarticulation	Auditory Models	
	Afternoon	first	1400-1540	Recognition of Telephone Speech	Text-to-Speech Synthesis I	Voice Source Characteristics	Speech Perception: Higher Order Processes I	Speech Production, Perception and Analysis
		second	1600-1740	Speaker-Independent Word Recognition	Text-to-Speech Synthesis II	Human Factors	Speech Perception: Higher Order Processes II	
WEDNESDAY	Morning	first	0830-1010	Continuous Speech Recognition I	Natural Language Processing and Speech Understanding I	Language Learning and Acquisition I	Speech Perception: Higher Order Processes III	Synthesis
		second	1030-1210	Continuous Speech Recognition II	Natural Language Processing and Speech Understanding II	Language Learning and Acquisition II	Prosody: the Phrase and Beyond I	
	Afternoon	first	1400-1540	Speaker Adaptation	Natural Language Processing and Speech Understanding III	Pronunciation Training	Prosody: the Phrase and Beyond II	
		second	1600-1740	PLENARY				
THURSDAY	Morning	first	0830-1010	Self Organizing Systems in ASR I	Synthesis I	Speech Perception: Phonetic Processes I	Speech Enhancement	Linguistic Phonetics
		second	1030-1210	Self Organizing Systems in ASR II	Synthesis II	Speech Perception: Phonetic Processes II	Speaker Recognition	
	Afternoon	first	1400-1540	Hidden Markov Models	Dialogue	Prosody and Phonology I	Speaking Styles I	Perception and Production
		second	1600-1740	ASR in Noise	Dialogues and Applications	Prosody and Phonology II	Speaking Styles II	
FRIDAY	Morning	first	0830-1010	Sampling of Speech Production I	Labelling of Speech I	Speech Production: Models & Theories I	Miscellaneous	Speech Recognition
		second	1030-1210	Sampling of Speech Production II	Labelling of Speech II	Speech Production: Models & Theories I	Speech Data Bases	
	Afternoon	first	1400-1540	Stochastic ASR	Analysis of Disfluencies in Speech I	Linguistic Phonetics: Reduction	Hearing/Speech Impaired I	Analysis/Systems
		second	1600-1740	Unit-Based ASR	Analysis of Disfluencies in Speech II	Language and Dialect Characterization	Hearing/Speech Impaired II	

表3.3-2　国別発表件数

米国	115	日本	89	カナダ	37	フランス	30
英国	36	ｵﾗﾝﾀﾞ	24	ｽｳｪｰﾃﾞﾝ	16	ドイツ	17
イタリー	14	中国	14	スペイン	10	ﾃﾞﾝﾏｰｸ	9
ﾌｨﾝﾗﾝﾄﾞ	6	台湾	4	ベルギー	4	韓国	4
インド	2	ﾉﾙｳｪｰ	2	その他9ｶ国	1		

分野ごとの研究発表の件数をまとめるとおよそ表3.3-3に示すようになる。この表からも分かるように、音声認識、理解、符合化、合成などのいわゆる工学的分野にとどまらず、音声知覚、生成、言語音声学などの分野の発表も数多くおこなわれていた。

表3.3-3　分野別の発表件数（概数）

（口頭発表）	
1）自動認識、理解（単語、連続音声、話者適応等）	75
2）符合化、合成	45
3）データベース、ラベリング、方言	15
4）発話解析（Speaking Style, Disfluencies） 対話解析	45
5）話者認識	5
6）韻律、音韻論、音声学	25
7）言語習得、獲得	15
8）音声知覚、生成	50
（ポスターセッション）	
1）分析、認識	58
2）合成	28
3）言語音声学	36
4）知覚、生成	61

第２部　音声認識･理解システム化技術

第１章　概　要

第1章　概　要

音声は、人が日頃気楽に使用しているため、機械にとっても音声認識・理解は比較的簡単に実現できると一部の人達には思われている。しかし、それは、人の音声認識・理解能力が人間の持つ極めて高度の知的処理機能に支えられているという事実に気が付いていないことによる。

ところが、現在の自動音声認識・理解技術は、一部実用化の段階に入っているとはいうものの、残念ながら、それを支えるべき知的処理技術はまだまだ未成熟である。そのために、適用できる対象はかなり制約があるのが実態である。

一方、現状の音声認識・理解技術を利用することにより大きな効果が期待できるアプリケーションがあっても、そのことをアプリケーション開発者が適切に判断ができない状態に置かれているのも事実である。現状技術がどの程度の能力を持っているのかを判断するための適切な情報がないためである。

これらのことが、音声認識・理解技術の普及の障害の1つとなっていると思われる。

また、音声認識・理解技術を利用できると判っても、いざそれをシステムに組み込もうとすると、大変な手間とノウハウの理解が必要となる。そのためにコストもかさみ、事実上実用化をあきらめざるをえない状態にもなっている。

音声認識・理解技術を普及させるためには、音声認識・理解技術自体の性能の一層の改良を進めることはもちろんのこと、高度の知的処理技術の開発、音声認識・理解技術をアプリケーション・システムに組み込むための支援ツールの整備、アプリケーションシステムの評価手法の開発、音声認識・理解部の組み込みインタフェース条件の標準化、音声認識・理解機能を現場で実際に使う人から見たユーザインタフェースの使用手順の標準化、等を進める必要がある。

以上の状況にかんがみ、第2部では、音声認識・理解技術組み込むための支援ツール、評価手法、インタフェース条件の標準化、ユーザインタフェースの使用手順の標準化など、応用開発支援関連の問題点に対応するためのいくつかの施策を提案する。

音声認識・理解技術の性能の高度化、高度知的処理技術の開発等の技術課題に関しては、第3部で述べることとする。

まず、第2章では、音声認識・理解技術をその構成要素の一部とする様々なアプリケーションシステムの分類を試みる。

分類の目的は、

（1）それぞれのタイプ（型）がどの程度の音声認識·理解技術を必要としているか、

（2）現状音声認識·理解技術がどのタイプまで対応可能か（どの様な応用に利用可能か）、

（3）現状音声認識·理解技術の対応の難しい各タイプについては、音声側の技術として、どの様な技術を今後開発する必要があるか、

など、その見通しを容易にする点にある。

このため、応用システムをインタフェース部（音声認識·理解関連技術）と問題解決部（応用目的の実現技術）の組み合わせと看做し、利用者の立場から望ましい両者の組み合せを考え、前者が共通となるように、後者をまとめて1つの型として分類した。

ここで提案したアプリケーションシステムの型はつぎの6種である。各型は音声認識·理解関連技術の特徴により分類されている。

A.同定型	特徴：信号レベルの同定、機械主導
	例：話者同定
B.項目入力型	特徴：単語認識、機械主導（人主導）、特定話者
	例：荷受けデータ入力システム
C.命令型	特徴：単語認識、機械主導（人主導）、不特定話者
	例：券売機
D.教育型	特徴：対話理解、話し言葉、機械主導
	例：英語ＣＡＩシステム
E.解釈型	特徴：対話理解、話し言葉、人主導
E-1　相談型	例：電話番号案内
E-2　解釈、計画型	例：旅行プラン作成システム
E-3　解釈、翻訳型	例：翻訳電話
F．ディクテーション型	特徴：文章認識、書き言葉、人主導
	例：ディクテーションシステム

第3章では、アプリケーション開発者が音声認識·理解技術をその構成要素の一部に簡単に組み込む際に利用できる、支援ツール（音声インタフェース開発支援ツール）の必要性を提言した。

また、その具体的例として、教育型の応用システム構築支援を想定し、音声対話アプリケ

ーション開発支援ツールとしての例を提案する。アプリケーションシステムに簡単に音声認識・理解機能を組み込むための記述形式や、アプリケーション（タスク）とユーザのモデルを記述する方法の試案も示す。

第４章では、現在利用が可能な音声認識・理解技術を実際にアプリケーションシステムに組み込む際に注意すべきノウハウ的留意事項を示した。

主なノウハウ的留意事項としては、

（１）話者条件、（２）発声形態、（３）語彙数、（４）文法、
（５）対話制御、（６）ユーザインタフェース、（７）収録系、
（８）評価法、（９）ハードウェア構成

等に関連した項目である。現在利用可能な製品や参考文献等を資料編に示した。

また、音声認識・理解技術を実際にアプリケーションシステムに組み込む際に、認識誤りができるだけ少なくなるように認識対象語彙セットを選択するための支援ツールの構築を提案している。これには、次の様な機能を検討した。

警告機能　与えられた認識用単語の中から，誤りの起こりやすい組み合せを検出し警告する。

選択機能　認識用単語を類義語の群毎にまとめて与え、全体として誤りの少ない組み合せを各群の中から選択、提案する。

提案機能　類義語辞書を用意し、与えられた認識用単語毎に類義語を検索し、全体として誤りの少ない組み合せとなるよう、各類義語の中から自動選択、提案する。

第５章では、利用者の立場から、音声認識・理解技術や、それを組み込んだアプリケーションシステムをどのように評価したら良いかについて、その試案を示した。ユーザ満足度、目標達成率、発声条件などの使いやすさ、認識・理解性能、導入による経済効果等について検討している。

第６章では、標準化について、検討を行う。現在、音声認識・理解装置は各メーカでそれぞれ独立に開発が進められているため、その使い方や、アプリケーションシステムへの組み込みインタフェース条件がバラバラになる恐れがある。このことは、最終利用者にとってメ

ーカが異なると操作方法が異なり、混乱を招くだけでなく、アプリケーションシステムの開発者にとっても設計上非常に面倒な問題となり、音声認識・理解技術の普及の妨げとなる。

本章では、この問題に対して、最終利用者にとっての操作方法の標準化と組み込みインタフェースの標準化の必要性を主張するとともに、関連分野の標準化動向について、先行しているアメリカの動きを中心に報告する。

最後に、第7章において、以上の各章の議論を踏まえ、今後の応用開発支援関連課題について検討を行う。

第２章　音声認識･理解応用システムの分類

第2章　音声認識·理解応用システムの分類

2.1　はじめに

音声認識·理解を実際に応用するためには、音声認識·理解技術を導入するメリットやデメリット、導入コストや現在の技術レベルを知ることが不可欠である。音声認識·理解技術は、現在のところ万能でなくコストも高く、現在の技術レベルの範囲で実現可能か、どのように構成するかをよく検討し導入に移らなければならない。本章では、このための音声認識·理解応用システムの分類に関し、音声認識·理解の機能、効果、システム特性、分類、制約、現状技術について音声認識·理解装置実用化の成功例を見ながら検討していく。考慮すべき導入時のメリット、デメリットには、開発者としてのものとエンドユーザに対する2種類がある。本章では次の問いに応えようとしている。

＜開発者の問い＞

○音声認識·理解で何ができるのか、何が便利なのか。

○音声認識·理解はどういうもので、どんな種類があるのか。

○どの分類に入り、実現性があるのか。

○どのくらいの性能が出そうか。

○どのように構成すればできるのか。

○どの程度のコストがかかるのか。

○問題点は何か。

○構成にあたって決めるべきことは何か。

○今まで実用化されているシステムの例と性格を知りたい。

○対象とする認識·理解装置のデータ利用性はあるのか。

＜エンドユーザーの問い＞

○音声認識·理解機能で便利になるのか。

○使用にあたっての制約は何か(登録、変換など)。

○どのくらいの性能が出るのか(認識·理解率、環境)。

○どういうものがすでに市販されているのか。

○いくらぐらいするのか。

2.2　音声認識·理解の機能

音声認識·理解の機能には、種々の機能があるが応用に関していえば、音声認識·理解でないといけない、音声認識·理解だからこそ実現する機能というものに注目する必要がある。容

易に代替手段が存在する機能を無理に音声で実現するのは合理的ではない。このような音声認識・理解独自の特長的機能としては、非接触制御機能、電話回線入力機能、多項目入力機能、音声対話機能の大きく分けて4つのものがある。以下個別に特長となる機能について説明する。

A. 非接触制御機能

非接触制御機能とは、手が使えない状態での装置制御を指している。通常、音声認識の代替制御手段として用いられているスイッチなどは手で制御されるので、手がふさがっている状態では制御不能に落ちいってしまう。このような場合には、やはり音声で制御することになる。音声認識・理解が不可欠なのは、この代替手段が使えない両手がふさがっている状態や、制御対象が手の届かない遠方の場合である。

例：・両手に鍋を持って、電子レンジのドアを開けたい。
　　・食肉市場の等級入力装置。手が汚れている状態で、肉の等級を入力する。

B. 電話回線入力機能

電話回線を経て音声で回線の向こう側の装置の制御を行なうものである。Aの非接触制御機能と良く似ているが不特定多数の音声による制御を受ける可能性があるため別の項目にしている。

例：・不特定多数の使用者の音声を認識して電話番号案内をする。
　　・自宅に電話してエアコンのスイッチを入れる。
　　・電話による各種情報検索

C. 多項目入力

一般に100とか1000とかいう多数の種類の中から項目を選択するのが多項目入力である。項目が多数になると、一つ一つに機能キーを割り当てることが、操作性、キーの占有する面積からできなくなる。このような場合にも、同時に全部を認識する必要がある場合と、本来構造のある項目から検索するので同時に認識する項目は少量ですむ場合がある。ここでは、前者を多項目横型検索、後者を多項目縦型検索と呼ぶ。多項目横型検索では、同時に全部を認識する必要があり高い認識精度が要求されるが、実際にはこの用途が非常に多い。

例：・工場の荷受けデータ管理
　　・物流センターの配送仕分けシステム

・歯科検査記録システム

多項目入力のうち、検索項目に構造がある場合の多項目縦型探索は、通常メニューなどで順に項目を選択することにより実現されている。しかし、メニューにより項目を検索するためには、メニューの構造をあらかじめ知る必要があり、これがわからないと検索できない欠点を持つ。音声認識により、極めて自然に並列に検索項目を入力することが可能となる。

D. 音声対話機能

前述の多項目縦型検索は、構造がある多項目の検索を行うものである。この場合、構造が一定であることが前提であるが、あらかじめ構造が規定できず使用者とのインタラクションによって構造が変化する場合がある。このような場合は、メニューを構成することは困難で音声対話により柔軟に対応することになる。

例：・地理案内システム
　　・航空券予約システム

E. 高速入力機能

音声入力は、入力手段への習熟が不要であり、入力速度が速く楽である。しかし、入力を長時間続けることが難しい欠点がある。

2.3 音声認識·理解による効果

音声入力の機能を生かしてシステムを構成することにより、システムとしての効果が生まれる。これは、実際に使用者に与えるメリットというよりむしろ導入者に対するメリットともいえる。　音声認識·理解応用システムを開発するためには、開発したシステムを導入してもらう必要があり、導入者のメリットを考えることが重要となる。ここでは、音声認識·理解による効果として、即時化、省力化、ユーザフレンドリー化、高信頼性化、低価格化、宣伝効果の6つの効果を考える。ただし、これらの6つの効果は、独立ではなく多かれ少なかれ相関がある。

A. 即時化

音声認識·理解により即時に入力が可能になる。今まで処理に時間がかかっていたデータ入力が即時に入力できるようになり生産性が向上する。

例：・工場の荷受けデータ入力の即時化

・物流センターの配送仕分けシステム

B. 省力化

導入することにより省力化、人手が省ける効果をいう。即時化と相関がある。実際には、導入者の最も大きな関心はこの点にあるといえる。省力化の程度には色々あって、そのオペレーションを受け持つ人自体を省力化できたり、手間を合理化することにより処理時間が早くなったり、作業量が低減したりして省力化が達成される。ただし、人の代行をさせる場合は、機械に対話機能などを持たせてヒューマンインタフェースを十分検討する必要がある。

例：・自動電話番号案内システム
・新幹線座席予約システム
・タクシー自動予約システム

C. ユーザフレンドリー化

ユーザフレンドリー化を行なうことで人間にとってより自然なインタフェースを実現し、また従来技術との差別化を行なう効果を指す。音声認識・理解は、本来人間が持つ通信手段なのでユーザに受け入れられやすいシステムを構成することができる。

例：・音声入力券売機
・ビデオ予約リモコン

D. 高信頼性化

手が使えないような応用において、文字などの入力手段を用いた場合は、データ入力に時間がかかり誤りが増える。これを音声入力で行うことでデータ入力速度を高く、かつ音声入出力によるインタラクティブな対話で確認を行うことで誤りを低減することができる。

例：・工場の荷受けデータ入力の即時化
・物流センターの配送仕分けシステム

E. 低価格化

入力を急ぐ応用などでは文字などによる入力手段を用いた場合、OCRなどの技術が要求され、また信頼性を上げるための機能を付加することで価格的にも高くなってしまう。このような応用においては、音声認識のインタラクティブな確認などの機能によって高信頼性ながらシステム全体の価格を低減することができる。

F. 宣伝効果

今まで使われていなかったところや、気が付かなかったところに音声認識·理解技術を用いることでシステムの宣伝効果を期待することができる。

2.4 音声認識·理解の形態と性能

音声認識·理解技術は現状では万能ではなく、高い認識率を達成するためには高速の処理系や記憶容量がいる。しかし用途によっては、話者の特定が可能で簡易な方法を使用でき、安価に装置化が実現できることがある。このため、システムを構築する場合、用途に設定できる条件をうまく利用しシステムを設計し実現することになる。ここでは、システムを構築するために必要となる音声認識·理解技術の条件、音声認識·理解の形態と性能について述べる。

A. 発声単位

使用者が認識装置に発声する発声の単位を指している。発声の単位には、大きく分けて次の6種類があり、順位が大きいほど難易度が高い。機能としては連続単語まで実用化されているが、連続単語モードでは性能が十分でなくあまり使われてないという調査結果もある。

・孤立単語発声

認識対象となる単語1単語を区切って発声する。認識は最も容易である。

・キーワード発声

認識対象となる単語を含む音声を発声する。検出すべき単語を探索するので計算量が多く、精度が落ちる。検出すべき単語が部分パターンの関係になっている場合に認識率が落ちる。

・連続単語発声

認識対象となる単語を連結して発声する。数字を連続して発声する場合などがこれに属する。

連結された単語数が既知の場合と未知の場合があり、未知の場合は非常に難しい。

・文節発声

日本語の文節で区切って発声する。”東京へ|いきます”という具合である。音声ワープロや自動翻訳電話プロジェクトでは、文節で区切った発声を要求している。

・読み上げ発声

文単位で発声された読み上げ調の音声のディクテーションを行う。読み上げ調であるが、自由文の入力を行うためタスクの限定などを行うことができず難易度は高い。

・自由発話

完全に自由な話しことばによる対話文の発声を指す。認識は極めて難しい。完全自由発話

の音声理解や、自動翻訳は相当将来にならないと難しい。

B. 発声形式

Aで述べた発声単位をどのように発声するかの問題である。装置によっては、色々な発声がなされ難易度は変わる。このため、発声のスタイルは1つの条件とみなすことができる。次の3つのタイプを考慮する必要があり、下にいくほど難易度が高い。また、雑発声に対しては現在の認識装置では性能が大きく劣化するのが実状である。

- 丁寧発声　明確に丁寧に発声する。
- 自然発声　普通に発声する。
- 雑発声　怒ったり、雑にして発声したり、ストレスをかけたりして発声する。

C. 話　者

どのような話者を対象にするか。大きく分けて、次の4つがあり、下にいくほど難しい。話者といっても声の音色の違いに加えて、話し方、方言の違いまであるので実際には不特定話者を認識するのは極めて難しい。現状では、話者に関しては認識語数との関係が強いので一概にどのレベルまで可能かを述べることは困難であるが、商品レベルでは100単語以上なら特定話者、それ以下なら不特定でも可能というところである。

- 特定話者
  ある決まった使用者の音声を認識する。その他の使用者については保証外となる。つまり、認識できるとも間違うともいえない。
- 複数特定話者
  複数の使用者がいて共通の認識対象単語がある場合に、その複数の使用者の音声を認識する。その他の使用者については保証外となる。つまり、認識できるとも間違うともいえない。
- 話者適応　(使用前に少し発声してもらえ、ある程度の間使用する)
  使用前に若干の音声を発声し、認識装置を使用者に適応化させてから認識する。このプロセスさえできれば誰の声でも認識できる。
- 不特定話者　(いきなり発声する)
  誰の声でも認識する。

D. 語　彙

どのような語彙を設定するかということである。あらかじめ決まった語彙なのか、使用者が

設定できるかなど分けると次の3つのタイプがある。下にいくほど難易度が高い。

・特定語彙

あらかじめ設計段階で認識対象語彙が決められているものである。数字の認識などは特定語彙である。単語単位の不特定話者音声認識では、多数の話者が該当単語を発声したデータベースを分析して標準パターンを作るのでほとんどの場合は特定語彙となる。

・可変語彙

使用者が語彙を設定できる。特定話者単語認識は、音声の登録時に音声パターンと語彙とを同時に設定する。音素、音節単位の認識を基本とするシステムでは文字列から認識語彙を設定できる。

・任意語彙

あらかじめ十分大きな辞書や文法を持っており、未知語(新出語)を容易に登録できる。このような機能を有するものを便宜的に任意語彙とする。

E. 語　数

認識語数は、応用に応じて設定される。不特定話者に対応する場合は、100単語以下が適当であろうし、特定話者なら100語以上可能だが登録が大変になる。このためシステムの制約と、認識性能との妥協点で設計することになる。

F. マイクロホン

使用する環境、使用形態、システムの構成によりマイクロホンの形態が決定される。音声を発声する口とマイクロホンの距離が近いほど認識性能を高くすることができるが、この距離の近い接話マイクロホンは話者を拘束するので使用者に好まれない。また、システムによっては、ワイヤレスの接話マイクロホンを用いるものもある。

G. 騒音環境

騒音環境は、それ自体音声の特徴パターンに歪みを与え認識性能を劣化させるばかりでなく、環境により音声パターンに発声変形を与える。これにより、認識性能は大きく劣化する。このためどのような環境において使用されるかで、認識方法、使用するマイクロホン、システム全体の設計を検討する必要がある。 大まかに分けて次のような騒音環境があり、下にいくほど難しいと考えれる。現在音声認識技術開発において対象とされているのは、家庭内、自動車内、駅の改札口、オフィスレベルなどである。

・家庭内環境

・オフィス環境

・自動車内(一般道路、高速道路、窓開け、オープンカー、雪道、トンネル)

・工場

・屋外(郊外、近郊、道路沿い、パチンコ店、駅のホーム…)

2.5 音声認識·理解応用システムの特性

音声認識·理解を応用してシステムを構成するにあたって考慮すべきこととして、インタフェースの関係、情報の向き、主導権、タスクの限定可能性がある。これによってシステムの構築の難易度が変わる。

A. インタフェースの関係

音声認識・理解システムを構築する場合、インタフェースの関係が重要となる。大きく分けて使用者の発声相手が機械か人間かということで発声者の挙動、発声の丁寧さが変わってくる。相手が人間だと早く発声するだろうし、発声の丁寧さも十分でない。相手が機械だと自分のいうことを理解してもらえないことを前提にするので発声は丁寧で明確になる。このため構築する機械がどのようなタイプで、どれぐらい発声の丁寧さが期待できるかを知る必要がある。インタフェースの関係としては、次の2つがある。

・(人⇔機械)型

・(人⇔機械⇔人)型

最初の(人⇔機械)型は通常の音声認識·理解システムである。(人⇔機械⇔人)型には認識符号化による通信や自動翻訳電話が含まれる。

B. 情報量

インタフェースの関係に加えて重要と思われるものとして情報の流れがある。一般に、情報は多い方から少ない方に流れるので、どちらが情報に多いかによって情報の流れが決まる。この情報の流れの方向によりシステムの設計がなされる。 話し手と相手の間での情報量の関係を次の3種類に分ける。

・ (人<機械)型

・ (人=機械)型

・ (人>機械)型

(人<機械)型には各種情報検索システムが入る。機械の情報を、音声応答による検索機能を用いて検索する。(人=機械)型には、音声指示、多項目入力がある。使用者は、機械にある項目をすべて知っていて音声入力によりどれかを指示する。(人>機械)型には、音声ワープロなどが入る。人間が頭の中に構成する文章を発声しそれを認識してタイプする。音声認識装置は一般に下にいくほど構築が難しい。1つの理由には発声の自由度が下にいくほど高いということがある。

C. 主導権

情報の流れに加えて、人間とシステムが対話を行う際の対話の主導権をどちらが有するかがシステム構成上大きなポイントとなる。主導権のとり方には次の3つのタイプがある。

・機械主導型
・人間主導型
・相互主導型

下に行くほど人間の発声の自由度が増すため難易度が高くなる。機械主導型では、人間は質問に答えるだけで良く、yes! no で答える質問だけを発する構成とすればシステムは更に簡単になる。人間主導の場合は、人間が多項目の音声や読み上げを行い認識する。認識語数はたいていの場合機械主導型に比べ大幅に増加する。相互主導型は、主導権が入れ替わりながら対話を進めるもので、話の途中で強制的に主導権が入れ替わることもあり難易度は極めて高い。

D. タスクの限定

音声認識·理解システムによっては、使用者の発声する音声のタスクが限定できる場合とできない場合がある(ここではタスクとは、発声内容の領域と文の難しさを同時に指している)。タスクが限定できる場合は音声認識·理解が易しくなり、限定できない場合は極めて難しい。

・タスク限定型
・タスクフリー型

情報検索、案内システムはタスク限定型であり、タスクを限定しながら音声対話を進めることができる。これに対し、音声ワープロなどは、タスクフリー型でいきなり何をいうか予想できないので認識は極めて難しいものになる。

2.6　音声認識・理解応用システムの分類

これまでに記述した音声認識の機能、音声認識の効果、音声認識・理解応用システムの特性をふまえ音声認識・理解応用システムの分類を行なう。分類においては、エキスパートシステムで用いられている分類の診断型、計画型、設計型、解釈型、制御型を参考にしている。

分類を表2.6-1に示す。

表2.6-1　音声認識・理解応用システムの分類

分　類	概要・機能	特　徴	応用例
A.　同定型	個人、 言語の同定	信号レベルの同定、 機械主導、 非接触、 即時	話者識別キー
B.　項目入力型	多項目入力	単語認識、 機械主導（人主導）、 特定話者、 非接触、 電話回線、 即時	荷受データ入力、 検査システム、 肉等級入力システム、 配送仕分システム、 歯科検査記録システム
C.　命令型	音声制御、 音声リモコン	単語認識、 機械主導（人主導）、 不特定話者、 非接触、 電話回線、 即時	音声認識電話、 音声認識券売機、 トーククレーン、 ビデオ予約リモコン、 ANSERシステム
D.　教育型	人間に知識を供与する （CAI）	対話理解、 話し言葉、 機械主導、 音声対話、 機械が人間を教育	英語CAIシステム
E.　解釈型	音声認識合成　＋ エキスパートシステム	対話理解、 話し言葉、 人主導	
E-1.　相談型	人間が理解するための 問い合わせ	音声対話、 人主導、 機械から提案	電話番号案内
E-2.　解釈、計画型	機械が理解しまとめる （申込み、予約）	音声対話、 機械主導、 機械がまとめる	旅行プラン作成システム
E-3.　変換、翻訳型	音声言語の翻訳、 認識符号化	対話理解、 人主導、 翻訳	翻訳電話
F.　ディクテーション型	文書作成のための入力	文章認識、 書き言葉、 人主導、 非接触、 高速入力、 タスクフリー	音声ワープロ、 ディクテーション システム

A. 同定型

音声を素材として音声に含まれる情報の同定を行なう。同定対象としては、現在のところ話者（個人）の同定がほとんどである。特長は、マイクに向かって発声するだけでたとえば装置の鍵を開けることができることであり、電話回線を通した話者同定への発展も考えられる。この機能により、今まで鍵や、指紋、アイパターンなどによっていた作業が即時化でき、よりユーザーフレンドリーな形での個人の同定が可能となる。

システム形態は、一般に機械主導でキーワードなどの発声を使用者に促し、音声を得て個人の同定、照合を行なうものが多い。キーワードが決まっている場合をテキスト依存型話者同定(識別、照合)、キーワードがなく、何かを喋れば同定するシステムをテキスト独立型と呼んでいる。現在のところ、1単語から数単語の決まった単語を丁寧に発声しテキスト依存で同定を行うものが実用に近いレベルで、テキスト独立、連続発声、自由発話音声などは研究段階である。また、電話回線を通した音声による話者の同定については、電話機自体のマイク特性や回線の特性のばらつきが大きいこと、個人性が電話帯域では十分に表現されないことからあまり積極的に研究されていない。その他、話者の同定のほかに言語の種別の同定なども試みられている。

B. 項目入力型

多項目にわたる項目指定を音声で行なうものである。荷受けデータ入力システム、配送仕分けシステム、歯科検査記録システムなど実用化された多くのシステムがあり省力化の観点からもっとも導入効果がある。このタイプは、本体システムが大きいため、音声認識部のコスト比率が低く、またオペレーターが限定されているので特定話者でも良く、更にワイヤレスマイクなども使えるため技術的制約を導入しやすく認識率の確保がやさしい。現在のところ、特定話者で数百語程度までのもの、不特定話者で数十語程度までのものが実用化されている。

C. 命令型

音声によるコマンド入力で各種機器、ロボットなどの装置制御に用いられる。特長は、非接触に声で指示するだけで機器を制御できるということである。システム形態は、比較的小規模のものでリモコンに変わるものから、電話回線を経た音声を認識するための大規模なものまで多様である。商品化されているものとしては、音声認識付電話機、音声入力券売機、ビデオ予約リモコン、アンサーシステムなどがある。認識性能は、使用する環境と装置にかけられるコストで決まるので、リモコン、電話などには特定話者で数十単語、不特定話者で10から20単語程度、電話回線を経たものでは不特定話者で数字と機能語が認識できる構成がとられている。しかし、他の入力手段に置き換えて得られるメリットよりも導入コストが高いのが現状で、どうしても音声でという音声の特長を活かした用途に使われてようやく実用化されている。課題

は、低価格化と認識性能である。認識性能としては、機器に装着されたマイクロホンを用いて、種々の使用環境で安定して高い認識率を達成することが課題である。現状のシステムではマイクロフォンを口から離したり、騒音および騒音に影響されることにより起こる音声の発声変形によって認識率は大きく劣化する。

D. 教育型

人間に知識を供与する型で、CAI、発音訓練、航空管制の訓練装置など各種教育装置がある。特長は、機械に教育すべき知識があって、教育に関する主導権があるということである。したがって、情報は、生徒の学習レベルを見ながら、ほぼ一方的にシステムから生徒に流れる。インタフェースとなる音声応答には、キーワード検出か対話文の認識·理解と音声出力が用いられる。現在、英会話のCAIシステムなどが開発されている。課題は、教育型エキスパートシステムと音声応答部にある。 音声応答部の課題は、対話音声の認識·理解と合成である。この分野では、まだ、ほとんど商品化がなされていないのが実状である。

E. 解釈型

音声認識、合成によるインタフェースと知識データベースを持つキスパートシステムが一体化したものをここでは総称して解釈型と呼ぶ。解釈型は、エキスパートシステムの分類にしたがって更に次の3種類に分類できる。

E-1. 相談型

人間が知識、情報を得るための情報検索装置である。機械側にある情報の問い合わせシステムであり、例えば電話番号案内システムがこの類にはいる。この型の場合、情報は機械にあって音声応答により使用者に伝達される。音声対話、キーワードスポッティングにより多項目の検索項目の入力を行なう。不特定話者の多項目音声認識、キーワードスポッティングの方法論も十分確立されておらず、全体システムのコストも非常に高くまだ実用化されていない。

E-2. 解釈・計画型

機械が理解しまとめを行なうタイプであり、申し込み、予約、旅行プラン作成システムがこのタイプに属す。 音声応答システムの音声応答部は相談型と同じ音声認識、ワードスポッティングの方式確立の問題を含んでいる。更に解釈、計画型では、使用者の要旨を汲んで適切なプランを提示する機構が必要であり課題が山積している。全体システムのコストも非常に高くまだ実用化の域に達していない。しかし、地理案内や航空券予約をタスクとして各国で現在盛んに研究されている。

E-3. 変換、翻訳型

音声の自動翻訳や、認識符号化により低ビットで伝送し音声合成により音声を復号するものである。音声の認識をして文字化し文字コードから音声を再生する型のものを指す。通信中に文字コードで翻訳を行なえば自動翻訳と呼ばれるシステムとなる。自動翻訳電話については、日本で1986年から産官学の共同プロジェクトとして発足し、1000語程度の国際会議申し込みタスクで翻訳ができることを示せた段階である。タスクをどううまく制限するか、使用語彙をどう制限するか、話者をどう制限するかが課題であり。、現在の認識·理解技術では、まだ実用化のレベルではない。

F. ディクテーション型

音声ワープロなどの文書作成のための認識システムを指す。バッチ処理のディクテーションシステムによる議事録作成システムなども考えられる。相談型や、解釈型、教育型は発声者の発声する単語文法をタスクを規定するということで制御できるが、文章作成型では発声者は任意の文章の入力を希望する。したがって、認識誤りを訂正するためにタスクやドメインの知識、文法拘束をうまく使うことができず、いかに音素の認識率を改善するかが課題となってくる。使い方、性能、ユーザーインタフェースをよほどうまく考える必要があり、現状での実用化は極めて困難である。

表2.6-1に各分類とこれまで述べた特徴、効果、機能などとの関係を示す。また、音声認識の性能仕様のところで上段に理想とする性能、下段に大まかな現在の性能を示した。現在、応用が行なわれているものは、ほとんどが制御命令型、項目入力型である。いずれも音声認識の必然性である非接触、多項目入力の特長を活かした用途展開となっている。多方、エキスパートシステムに関係する教育型、相談型、解釈型、計画型、変換、翻訳型はまだまだ研究段階であるといえる。

2.7 音声認識·理解システムの現状

ここでは、実際に商品化され稼働している音声認識·理解応用装置について分析し、これらを通して応用システム構築の方法論、課題を明確にしていく。表2.7-1は平成3年度報告書で開発事例として紹介した装置について、これまでに述べた分類の各要素により整理を行ったものである。以下ではこの中から更に抜粋して分析をする。

A. 配送仕分けシステム

配送センターでベルトコンベヤー上に流れてくる荷物の送り先の振り分けを音声で指示する。送り先ラベルが見えない場合は、手で向きを変える。音声で入力するため処理速度が早く、両手が空いているために荷物の向きを変えるのも容易である。用いられているのは孤立単語の特

定話者音声認識であるが、使用環境が工場のため極めて高騒音下での高認識率が要求される。ここでは、ヘッドセットマイクとワイヤレスマイクを用いて耐騒音性を高めている。

B. 肉等等級入力システム

肉の等級を入れるためには、リフトに吊られて流れてくる牛肉を回しながら等級を人間が判断する必要がある。肉を回す際には手を使うが、手が汚れていることから等級の記録に手を使えず、ここに音声認識を使っている。手が使えない場面での利用という利点を最大限に利用している。認識環境は、配送仕分けと同じで工場内の認識であるが、接話型マイクロホンとワイヤレスで耐騒音性機動性を高めている。

C. 歯科検査記録システム

歯科検査を行なう場合、歯科医が歯の番号と検査結果を口で言って、助手が書き留める形式であったが、この助手の作業を自動化するものである。特定話者で孤立発声だが手が使えない用途への応用で効果的である。診察室なので騒音レベルも低い。

D. 音声制御クレーン

クレーンを操作する時、操作を音声認識により自動で行なうもの。危険を伴うが、認識結果を音声合成による応答で必ず確認することで回避している。遠方の操作を音声認識で行なうので作業員の省力化が実現する。特定話者で孤立と連続の発声が可能で、接話マイクとワイヤレスマイクで耐騒音性機動性を高めている。

E. 電話回線サービス

電話回線を経た数字と機能語音声を認識する。省力化、自動化という観点から効果は大きい。他の通信機器との連携で更に効果が上がっている。電話機の多品種化が進み音声品質にバラツキが増え認識率が低下してきている。音声認識による電話番号案内や通信販売などでニーズが大きい。

F. 音声入力券売機

試作機の性能評価が切符売場で行なわれた。通常、行き先の絵を見て値段を調べ値段通りの金額を自動販売機に投入し、値段のボタンを押すが、行き先の名前を入力するだけで金額を知ることができる。騒音が高く不特定話者のため技術的課題が多い。

G. ビデオ予約リモコン

現在まで繁雑と言われてきたビデオの予約を音声認識で行なったものである。録画時刻と時

間を言う。不特定話者単語音声認識を用いている。約380人の音声を分析して標準パターンを作成。　ユーザーインタフェースの工夫もあり評判は良好。

H.　ハンドフリー音声認識電話

音声認識により電話を発信する。ダイヤルする手間を省く効果がある。必然性は少し低いが、ユーザーインタフェースとして優れる。特定話者孤立発声のものがほとんどである。語数が少ないので認識率は比較的高いが電話に入れるにはまだ少しコストが高い。

表2.7-1　応用例の分類　(1/2)

	分類	機能					効果			
		非接触	電話回線	多項目	音声対話	高速入力	即時化	省力化	U/F改善	高信頼性
·配送仕分けシステム	項目入力	○		○		○	○	○	○	○
·肉等階級入力システム	項目入力	○		○		○	○	○	○	○
·歯科検査記録システム	項目入力	○		○		○	○	○	○	○
·麻酔記録装置	項目入力	○		○		○	○	○	○	○
·荷受け業務音声入力	項目入力	○		○		○	○	○	○	○
·検査結果入力システム	項目入力	○		○		○		○	○	○
·生産管理システム	項目入力	○		○		○		○	○	○
·音声制御クレーン	制御·命令型	○				○	○	○		
·電話回線サービス	解釈·計画型		○		○	○	○	○	○	
·タクシー電話自動受付	項目入力		○				○	○		
·電話オペレータサービス	解釈·計画型		○				○	○		
·スペースシャトル	制御·命令型	○		○			○	○		
·音声入力券売機	項目入力			○		○	○		○	
·音声入力エレベータ	制御·命令型	○					○		○	
·ビデオ予約リモコン	制御·命令型			○			○		○	
·ハンドフリー音声認識電話	制御·命令型	○		○		○	○		○	
·音声ワープロ	文書作成型			○		○	○		○	
·英会話CAIシステム	教育型			○	○		○	○	○	

表２.7-１　応用例の分類（2/2）

	形態と性能　(上段:要求仕様(想定)、下段:現状)							システム特性			
	A 発声単位	B 発声形式	C 話者	D 語彙	E 語数	F マイク	G 雑音環境	A I/F関係	B 情報量	C 主導権	D タスク
·配送仕分けシステム	連続単語 孤立単語	自然発声 丁寧発声	不特定 特定	可変 可変	500 150	タイピン型 接話型	工場 工場	人⇔機械	人>機械	機械主導	限定
·肉等階級入力システム	連続単語 孤立単語	自然発声 丁寧発声	不特定 特定	可変 可変	100 18	タイピン型 接話型	工場 工場	人⇔機械	人>機械	機械主導	限定
·歯科検査記録システム	連続単語 孤立単語	自然発声 丁寧発声	不特定 特定	可変 可変	500 -	タイピン型 接話型	事務所 事務所	人⇔機械	人>機械	機械主導	限定
·麻酔記録装置	連続単語 孤立単語	自然発声 丁寧発声	不特定 特定	可変 可変	500 -	タイピン型 接話型	事務所 事務所	人⇔機械	人>機械	機械主導	限定
·荷受け業務音声入力	連続単語 孤立単語	自然発声 丁寧発声	不特定 特定	可変 可変	100 20	タイピン型 接話型	工場 工場	人⇔機械	人>機械	機械主導	限定
·検査結果入力システム	連続単語 孤立単語	自然発声 丁寧発声	不特定 特定	可変 可変	500 100	タイピン型 接話型	工場 工場	人⇔機械	人>機械	機械主導	限定
·生産管理システム	連続単語 孤立単語	自然発声 丁寧発声	不特定 特定	可変 可変	500 100	タイピン型 接話型	工場 工場	人⇔機械	人>機械	機械主導	限定
·音声制御クレーン	離散単語 孤立単語	自然発声 丁寧発声	不特定 特定	可変 可変	100 21	タイピン型 接話型	工場 工場	人⇔機械	人>機械	人主導	限定
·電話回線サービス	連続単語 孤立単語	自然発声 丁寧発声	不特定 不特定	可変 固定	100 16	ハンドセット スピーカホン	電話回線 電話回線	人⇔機械	人<機械	機械主導	限定
·タクシー電話自動受付	連続単語 孤立単語	自然発声 丁寧発声	不特定 不特定	可変 固定	50 10	ハンドセット スピーカホン	電話回線 電話回線	人⇔機械 ⇔人	人>機械 人<機械	機械主導	限定
·電話オペレータサービス	連続単語 孤立単語	自然発声 丁寧発声	不特定 不特定	可変 固定	100 5	ハンドセット スピーカホン	電話回線 電話回線	人⇔機械	人<機械	機械主導	限定
·スペースシャトル	連続単語 孤立単語	自然発声 丁寧発声	不特定 特定	可変 可変	100 41	接話型 接話型	日常 事務所	人⇔機械	人>機械	機械主導	限定
·音声入力券売機	孤立単語 孤立単語	自然発声 丁寧発声	不特定 不特定	可変 固定	26 26	機器装着型 機器装着型	工場 工場	人⇔機械	人>機械	人主導	限定
·音声入力エレベータ	孤立単語 孤立単語	自然発声 丁寧発声	不特定 不特定	可変 固定	50 5	機器装着型 機器装着型	日常 事務所	人⇔機械	人>機械	人主導	限定
·ビデオ予約リモコン	連続単語 孤立単語	自然発声 丁寧発声	不特定 不特定	可変 固定	50 15	機器装着型 機器装着型	日常 事務所	人⇔機械	人>機械	機械主導	限定
·ハンドフリー音声認識電話	連続単語 孤立単語	自然発声 丁寧発声	不特定 特定	可変 固定	100 20	ハンドセット スピーカホン	自動車内 自動車内	人⇔機械	人>機械	機械主導	限定
·音声ワープロ	文章 文節	自然発声 丁寧発声	不特定 特定	可変 可変	任意文章 8万語	タイピン,卓上 接話型	日常 事務所	人⇔機械	人>機械	人主導	フリー
·英会話CAIシステム	対話 文節	自然発声 丁寧発声	不特定 不特定	可変 可変	任意対話 4万語	タイピン,卓上 接話型	日常 事務所	人⇔機械	人<機械	人主導	フリー 限定

I.　音声ワープロ

音声ワープロは、音声により任意の文章を入力できる。キーボードに不慣れな使用者がワープロで文書を作成するために有効な機能である。音声ワープロのようなタスクなどによらない認識は、特定話者でも極めて難しい。現在開発されているのは、特定話者、文節発声のものだが基本性能がまだまだ十分でなく今後の発展が期待される。

2.8 おわりに

音声認識･理解応用システムの特徴、機能などを整理し分類をした。また、実際に商品化されているシステムについて若干ふれ、狙いや課題の分析を行なった。現在のところ、不特定話者で連続認識･理解の可能な商品を実現するには技術が十分でなく、より高い性能を追求するためには、高速の処理系、大量の記憶容量が必要となってくる。このため、設計に際しては、音声認識･理解が万能でないということを考慮し、当該応用で許される条件を最大限に利用し認識率を高めコストを削減する必要がある。また、認識･理解性能ばかりでなく、ユーザインタフェースとして、特に認識誤りを起こした場合の処理や、他のメディア(文字や、ジェスチャー)との協調を考えることが重要な課題である。

第３章　音声対話アプリケーション
開発支援ツールの提案

第3章　音声対話アプリケーション開発支援ツールの提案

3.1　音声対話アプリケーション開発支援ツールの必要性

音声認識･理解システムの利用拡大を考える上で、認識率や耐騒音性など音声認識技術の性能向上に加えて、一般のシステム開発者やユーザが現状の音声認識技術を用いて、実用的なアプリケーションを簡単に開発できるような環境を提供することが重要であると考えられる。特に、音声アプリケーションには多かれ少なかれ必ず存在するシステムとユーザ間の対話制御に関しては、簡単なツールを提供しない限り、数多くの音声アプリケーションがユーザの手によって開発されていく状況を期待することは難しい。もちろん、自然で自由な発話による音声対話など高度で複雑なシステムの実現は非常に難しく、今後のこの分野の研究の進展を待たねばならない。しかしながら、アプリケーション分野のなかには、第2部 第2章で述べた音声認識･理解応用システムの分類の内、項目入力型、命令型、教育型、解釈・計画型など、機械主導型の非常に限定された会話でその目的が達成でき、その対話の流れも状態遷移で記述できるような比較的簡単なものが数多くある。そこで、このようなアプリケーション用に、音声認識や言語処理の研究開発者ではなく、アプリケーション開発者やユーザが、対話の制御を比較的簡単に定義できるツール、「音声対話アプリケーションビルダー」を提案する。

音声対話アプリケーションの開発に一般的な手順が確立されているわけではないが、連続音声認識による対話システムを想定すると、そのアプリケーション開発には次のようなステップが考えられる。

（1）アプリケーションの選定
（2）対話例の採取
（3）対話の流れの定義
（4）語彙、文法の定義
（5）システム全体の開発、他のインタフェースとの統合
（6）テストと評価

音声対話アプリケーションビルダーは、これらすべてのステップにおいて、アプリケーション開発者を支援するものであることが望ましいが、ここでは（3）、（4）、（5）を支援するものを考える。（6）については、第4章で述べられている。

3.2　音声対話アプリケーションの例

3.2.1　日本語会話CAIシステムの概要

音声認識を用いた対話的なアプリケーションで、比較的対話の内容が簡単で実用化が近いものの例として、外国語会話の訓練システムが考えられる。すでに、音声認識やビデオ動画を用いて、インタラクティブに訓練を進める英会話CAIシステム［山本 91］［山本 92］があるが、ここでは、このシステムの考え方に基づき、外国人のための日本語会話CAIシステムを考え、題材として取り上げる。日本語会話訓練の中には、当然単語、熟語、構文、文法などの訓練も含まれるが、ここでは音声認識、録音音声出力、ビデオ動画などを用いて、学習者がシステムと会話しているかのような体験ができるシミュレーション会話を音声対話の対象とする。

日本語会話CAIシステムでは、学習者がシミュレーション会話に入る前に、その場面の動画ビデオを見ることにより、場面の状況（登場人物の名前、会話の目的、約束の日時、内容など）を理解する。その後、基本練習に入り、その場面で用いられた会話文の表現、構文などを練習し、あとのシミュレーション会話で、これらの表現が使えるようにする。表3.2-1に、電話による食事の約束を題材とした、システムと学習者の会話の例を示す。

なお、電話の取次に関しては［田路 91］に、かなり自由な発話を許す対話制御の方法が示されている。

表3.2-1　シミュレーション会話の例

システム：はい、ＡＢＣ商事です。
学習者　：私、イロハ貿易の田中と申しますが、営業部の鈴木一郎さんいらっしゃいますか？
システム：はい、ただいまつなぎますのでしばらくお待ちください。
システム：もしもし、鈴木ですが。
学習者　：イロハ貿易の田中です。ご無沙汰しております。
システム：こちらこそ、ご無沙汰しております。お元気ですか。
学習者　：はい、おかげさまで元気にしております。ところで、私は今大阪におりますので、もし、今晩お時間がありましたら、お食事でもご一緒にいかがですか。
システム：それはいいですね。では、駅前ホテルのロビーで７時にお会いするというのはいかがでしょうか。
学習者　：結構です。では、今晩七時に駅前ホテルのロビーですね。
システム：そうです。では、のちほど。
学習者　：はい。失礼いたします。

3.2.2　日本語会話CAIシステムの構成

前述のような、日本語会話CAIシステムの構成例を図3.2-1に示し、以下にそれぞれのモジュールの機能を述べる。

（1）入力インタフェース部

入力インタフェース部は、音声入力、キーボード入力、メニュー入力などの入力を扱う。音声認識とキーボードによる文入力は、入力文解析部に送られる。

（2）入力文分析部

入力文解析部では、文のマッチングや文法的な解析を行って、意味的内容をタスクモデルに送る。

（3）タスクモデル

タスクモデルは日本語会話CAIシステムの場合、教材知識と教授知識からなる。タスクモデルの詳細については、3.4で述べる。

（4）ユーザモデル

ユーザモデルは、日本語会話CAIシステムの場合学習者モデルである。これについても、3.4で述べる。

（5）出力インタフェース部

出力インタフェース部は、音声、テキスト、メニュー、静止画、動画などの出力を扱う。各出力の内容、タイミングはタスクモデル部から指示が出される。

（6）アプリケーションビルダー

アプリケーションビルダーは、上記（1）から（5）までのモジュールの内容を、開発者が記述するための支援ツールで、詳細は以下に述べる。

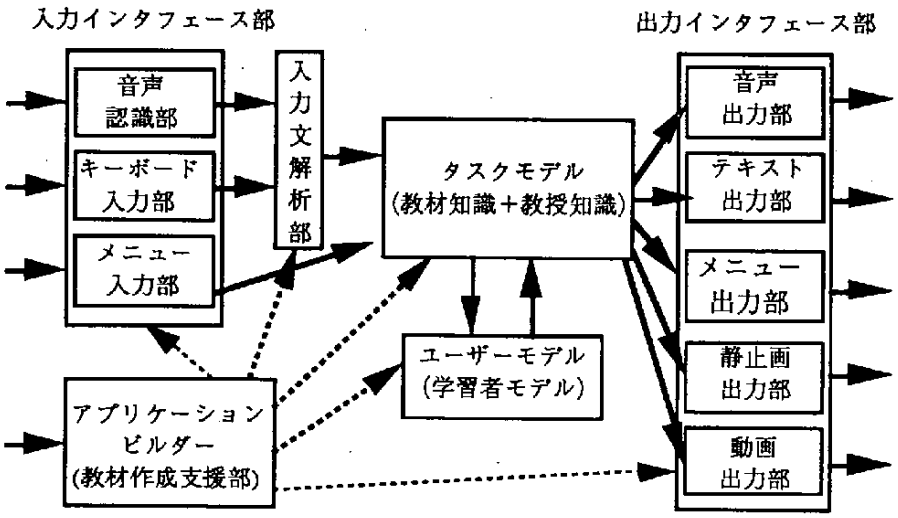


図3.2-1　日本語会話CAIシステムの構成

3.3　音声対話アプリケーション開発支援ツールの構成例

3.3.1　カード型エディター

対話の流れの定義、語彙・文法の定義、システム全体の開発、他のインタフェースとの統合などの開発支援ツールとして、カード型のアプリケーション記述エディターを考える。カード型エディターは、システムカード、学習者カード、文法カード、スクリプトカードなどから構成される。

例として取り上げた日本語会話CAIシステムの対話は、システム側と学習者側の発話が交互に行なわれることにより進行する。そこで、システム側の発話とそれに関連したデータを記述したシステムカードと、学習者側(ユーザ側)の発話と関連データを記述した学習者カード(ユーザカード)の2種類のカードを交互に実行させていくことにより、アプリケーションが定義できるようにする。　図3.3-1にシステムカードの例、図3.3-2にそれに続く学習者カードの例を示す。また、学習者の発話の認識に必要な文法を記述するための文法カードを図3.3-3に示す。また、システムカードと学習者カードを直接リンクさせてアプリケーションを定義することもできるが、発話の順序などに多少の自由度を与えるために、その場面の定形的なストーリーを記述するためのスクリプトカード（図3.3-4）を用意している。スクリプトについては、3.4で詳しく述べる。

システムカード	
カードID: S010	
行動主体：システム	役割：電話受付者
発話：「はい、ＡＢＣ商事です。」	
ビデオ動画表示： 装置：Dev1　　表示位置：Window002-3 Start Frame:A10081　　End Frame: A10125	
文字表示： 文字種：明朝12　　表示位置：Window001-1	
アクション： SCRIPT SSC010 着信発話 done goto CARD LSC010	

図3.3-1　システムカード

学習者カード	
カードID: L020	
行動主体：学習者	役割：田中太郎
音声認識： 装置：Dev2　モデル：男性15　しきい値：650 文法：G010	
アクション： IF < Score 650 THEN goto CARD S1010 IF Recog is *I AM TANAKA* and *WANT SUZUKI* THEN SCRIPT LSC010 自己紹介 done and SCRIPT LSC010 相手指定 done and goto CARD SSC010 IF Recog is *WANT SUZUKI* THEN SCRIPT SC010 相手指定 done and goto CARD SSC010 goto CARD SSC010	

図3.3-2　学習者カード

3.3.2 システムカード

図3.3-1のシステムカードは、主にシステム側の発話とそれに伴うビデオ動画を表示するためのカードである。「行動主体：」の欄はシステムが電話受付者の役割を演じていることを示している。「発話：」の欄には発話内容の「はい、ＡＢＣ商事です。」を示している。「アクション：」欄は、これらの出力後のシステムの行動を記述するが、ここではシステム側スクリプトカードの「着信発話」の項目にフラッグを立てた後、無条件に学習者側スクリプトカードに進むことを示している。

3.3.3 学習者カード（ユーザカード）

図3.3-2の学習者カード（ユーザカード）は、学習者側の音声発話を認識し、それに対するアクションを決めるためのカードである。「音声認識：」の欄では、認識装置のデバイス名、認識モデル名、認識スコアのしきい値、文法を記述した文法カードの ID 番号などを指定する。ここで用いる文法カードG010 は図3.3-3に示されている。「アクション：」欄では、学習者発話の音声認識結果をもとにシステム側の次の行動を決める。表現形式は、ルールベースシステム的な IF, THEN, ELSE 方式を用いている。IF で示された条件部を上から順番に照合し、マッチすればそれに対する THEN 部を実行する。このカードでは、1番目の IF - THEN 部で、認識スコアがしきい値より低ければ、システムカード S1010 に進み、学習者に再発話を要求する。2番目の IF - THEN 部では、認識結果が文法カード G010 の *I AM TANAKA* と *WANT SUZUKI* で定義される文であれば、学習者側のスクリプトカード LSC010 の「自己紹介」と「相手指定」の項目の実行済欄に「done」を記入した後、システム側のスクリプトカード SSC010 に進むことを示している。以下省略するがこのように、学習者の発話の認識結果に基づきシステムの次の行動を決定している。

3.3.4 文法カード

図3.3-3の文法カードでは、学習者の発話を認識するための文法を記述する。ここでは、文法として有限状態オートマトンを用い、これを簡便のために図に示したような表記で定義する。この文法は、自分の名前を言う *I AM TANAKA* と、話したい相手の名を言う *WANT SUZUKI* から構成されているが、それぞれ複数の表現が許容されるようにしている。

3.3.5 スクリプトカード

図3.3-4のスクリプトカードは、その場面の定形的なストーリーを定義するためのカードである。このスクリプトカードは、「電話をかけて食事の約束をする」場面を構成する学習者

側の行動の項目とその内容、それに対応した学習者カードの ID 、そのカードの実行を確認するフラッグなどを記述している。この場面の対話の制御は、おおむねこのスクリプトカードの項目を上から順に実行することによって実現されるが、細かい制御はシステムカードや学習者カードのアクション：の欄に記述されており、時にはある項目を飛ばしたり、後に実行したりすることも可能である。他に、システム側のスクリプトカードもある。

文法カード

カードID: G010

文法：

S -> (*I AM TANAKA*) *WANT SUZUKI*

I AM TANAKA
-> （わたし（は）｜わたくし（は））
（ABC 商事の）
田中　（一郎）｛です｜と　申し　ます｝

WANT SUZUKI
-> （営業部　（の））鈴木　（一郎）（さん｜様）
｛（を）お願い　します
｜（は）｛いらっしゃい｜おられ｝　ます　か｝
.....

図３.３-３　文法カード

スクリプトカード

カードID: LSC010

スクリプト名：電話をかけて食事の約束をする

項目	内容	カード	実行済
自己紹介	自分の名と会社名を言う	L020	done
相手指定	話したい相手の名を言う	L040	done
相手確認	電話に出た相手を確認する	L060	
開始挨拶	はじめの挨拶をする	L080	
都合確認	相手の都合を確認する	L100	
日時場所	会う日時と場所を決める	L120	
終了挨拶	終わりの挨拶をする	L140	

アクション：

goto CARD LSC020

図３.３-４　スクリプトカード

３.３.６　実行環境と実行例

カード型エディターで記述した対話制御を実行するためには、実行環境を必要とする。図３.３-５にこの実行環境の例を示す。カードスタックは日本語会話CAIのアプリケーションを定義した、複数のシステムカード、学習者カード、文法カード、スクリプトカードなどからなる。カードインタプリターはメインプログラムの指示にしたがって、これらのカードを必要な順番に読み込み、内容をメインプログラムに渡して行く。メインプログラムは必要に応じて、デバイスライブラリーを用いて入出力デバイスを操作する。

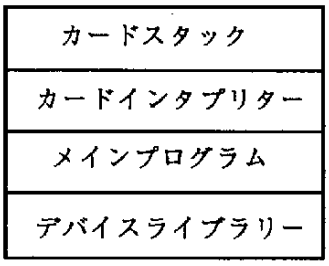


図３．３-５　実行環境

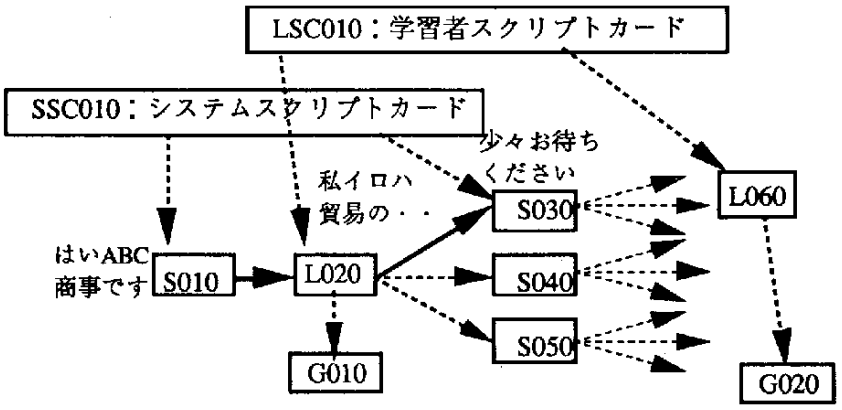


図３．３-６　実行制御例

次に、前述のカードを実行したときの流れの一部を図３．３-６に示し、その詳細を以下に説明する。

（１）まず、アプリケーションが開始されるとシステムスクリプトカード SSC010 が、システムカード S010 を指定して、システム側が「はい、ABC 商事です。」と発話した後、学習者スクリプトカードに主導権を渡す。

（２）学習者スクリプトカードは、デフォルトとしてまだ実行済みのフラッグの立っていない、最上段の項目、「自己紹介」のカード L020 を指定して学習者の発話を待つ。カード L020 は、文法カード G010 の文法をデフォルトとして指定する。学習者が「私、イロハ貿易の・・・」と発話すると、マッチングした発話内容のカテゴリーに従って、学習者スクリプトカードでその項目に実行済みのフラッグを立てた後、システムスクリプトカードに主導権を渡す。

（３）システムスクリプトカードは、つぎのシステムカード S030 を指定して、「少々お待ちください・・・」と発話した後、学習者スクリプトカードに主導権を渡す。

（４）学習者スクリプトカードは、同様にして学習者カード L060 を指定して学習者の発話を待つ。

（5）以下、同様にして対話を進めていく。

このように、対話の目的、発話の順序、語彙、文法などに比較的強い制限を設けて良い場合には、以上に述べたアプリケーションビルダーを用いることによって比較的簡単に、音声対話の制御を記述することができる。ここで示したアプリケーションビルダーは、あくまでも例であって、実現方法は他にも数多く考えられるが、今後の音声認識·理解システムの普及、応用拡大を促進するために、このようなアプリケーションビルダーが開発されユーザに提供されることが望まれる。

参考文献

1. [山本 91] 山本秀樹他、"音声入力可能な英会話用知的ＣＡＩシステム"、電子情報通信学会、教育工学研究会資料、ET91-107、57-62、1991
2. [山本 92] 山本秀樹他、"環境型語学用知的ＣＡＩシステム"、情報処理学会、コンピュータと教育研究会資料、20-5、1-8、1992
3. [田路 91] 田路龍太郎他、"電話取次対話における対話主導権の動的制御"、電子情報通信学会論文誌、D-II Vol.J74-D-II、No.12、1639-1950、1991

3.4　タスク/ユーザモデルの構成例

3.4.1　はじめに

本節では知的インタフェースの構成要素であるタスクモデルとユーザモデルについて、これまでの研究成果を簡単に紹介し、前節の例を使って具体的に説明する。

3.4.2　タスクモデル

すぐれたインタフェースを提供するためには、対象となるタスクを正確に記述したタスクモデルが必要となる。しかしながら、タスクをどのような階層構造に分解するか、タスクで用いる基本的なオペレータにはどのようなものがあるのかなどはタスクに強く依存し、一般化が困難である。このためタスク分析の手法はまだ確立されているとはいいがたい。ここでは、比較的小規模なタスクのモデル化に利用できそうなスクリプトとMOPについて説明する。

（1）スクリプト

スクリプトは、人間が日常的におこなう定型化された状況を事象の因果関係による連鎖とい

う形式で表現する記憶構造のことである［Schank 77］。スクリプトは自然言語理解において文脈的な知識を扱うためにSchank らによって提案された。スクリプトはその状況の中に登場する物や役割、そのスクリプトが適用されるための条件などとともに、シーンと呼ばれる事象の因果関係にそった連鎖から構成される。たとえば、有名なレストラン・スクリプトでは、登場物として、テーブル、メニュー、料理、お金、領収書などが、役割として、客、ウェイター、コック、会計係、店の所有者などを記述する。シーンとしては「店に入る」、「注文する」、「食べる」、「お金を払う」、「店を出る」などの事象の連鎖を記述する。それぞれのシーンは概念依存表現［Schank 75］で表現されている。このようなスクリプトを用意することによって、言語表現では明示的に表現されていないことについても推論をおこなうことができる。たとえば、「太郎はレストランへ行って、チキンを食べた。」という1文について「太郎は何を注文したのか?」、「太郎はお金を誰に払ったか?」といった質問に答えることができる。人間は多くのスクリプトを記憶しており、状況に応じて適切なスクリプトを想起して言語を理解しているというのがSchank らの主張である。

タスクを記述するのにスクリプトを用いれば、ユーザが処理の手順をある程度飛ばしても、そのギャップを埋めて適切に対処できる。特に音声インタフェースでは、ユーザの次発話を予測するための指針としてスクリプトを利用することができる。ただし、スクリプトには、あらかじめ用意されたシーンの連鎖をはずれた場合に対処ができないという欠点があるので、タスクが複雑になるとすべての場合を網羅したスクリプトを用意するのは困難である。

（2）MOP

スクリプトは個別の状況における事象の連鎖を記述したものであった。しかし、世の中にはよく似ているが、少しだけ違うような状況が多数存在する。言いかえれば、多くの状況で共通に出現するような事象が多数存在する。スクリプトによる表現では、個々のスクリプト同士は関係付けられていないために、よく似た状況のとり違え、たとえば、「歯科医に行く」と「医者に行く」をとり違えるという現象が説明できない。また、同じようなシーンをスクリプトごとに記述しなければならない。このような問題を解決するためにSchank は、スクリプトをより小さい単位に分解し、それらを動的に組み合わせるMOP（Memory Organization Packet）と呼ばれる記憶構造を提案した［Schank 82］。MOP の枠組では、たとえば、レストラン・スクリプトの「お金を払う」というシーンは「お金を払う」というMOP として表現され、このMOP は、「レストランで食事をする」という状況以外でも使用される。

3.4.3 ユーザモデル

ユーザモデルはシステムが持つユーザの知識と定義することができる。知的インタフェース

はユーザに応じて応答を変えることが必要である。たとえば、ユーザが初心者なのか、熟練者なのかによって適切な応答は違ってくるだろう。ここでは、Kass のユーザモデルの分類を紹介する［Kass 89］。この分類はITS（Intelligent Tutoring System）のユーザモデルを対象とした分類である。ITS の目標はユーザが専門家の知識を獲得することであり、システムは専門家の知識を参照できるから、ユーザモデルは専門家モデルと関係づけながら作るのが有効である。ユーザモデルと専門家モデルの関係は以下のように分類できる。

（1）Overlay Modeling

もっとも単純なモデルで、専門家の知識の部分集合としてユーザの知識を記述する。「ユーザは専門家の知識の一部を持っていないために専門家とは違うパフォーマンスを示すのだ」という考え方である。このモデルはWUSOR-II［Carr 77］で使われている。WUSOR-II は"Hunt theWumpus"というゲームのアドバイザであり、どのように迷路を探索するか、どのように危険に対処するか、行動の決定のための確率的な推論方法などをユーザに教える。
WUSOR-II がoverlaymodelingでうまくいくのは、以下の3 つの理由による。

・システムのゴールが専門家の知識を教えることだけである

・専門家モデルが真に専門家の知識である

・専門家モデルが非常に単純である

しかし、診断の戦略に関する知識など、専門家の知識がもっと深く、ユーザとの共有知識の認識がむずかしい場合にはoverlay model は適さない。また、定義から明らかなように、ユーザが専門家とは違った知識を持つことが意味を持つような場合にも適さない。ユーザが間違った戦略を信じている場合、overlay model はそれを指摘することはできるが、なぜそれが間違っているのかを説明することができない。

（2）Differential Modeling

ユーザと専門家の知識を比較するかわりに動作を比較する。専門家と同じ動作が得られれば十分という立場から、専門家の知識を専門家と同じ動作を得るのに必要な知識とユーザが知らなくてもよい知識に分ける。システムは、ユーザの実際の動作と専門家モデルでシミュレートした動作を比較して、ユーザが同じところでいつも失敗するのがわかると、それに対応する知識がユーザに欠落していると判断する。differential model は、ユーザの初期知識の範囲を決めつけていたというoverlay model の問題は解消したが、ユーザの知識が専門家の知識の部分集合である点は変わらず、overlay model の問題のほとんどを残している。

（3）Perturbation Modeling

ユーザは専門家が持たない知識を持つことができる。多くのITS がこのモデルを採用している。専門家の知識の他に、ユーザが誤解して持っている知識(バグ) を記述することによって実現するのが一般的である。たとえば、DEBUGGY システム［Brown 78］では、引き算の正しいやり方を手続きネットワークで記述し、その上にユーザの間違いを表現する手続きを付加する。perturbationmodel では、ユーザの失敗が知識の欠落による場合と間違った知識による場合があるため、ユーザモデルを作る際の探索空間が大きくなる。このため、探索空間を絞り込むための機構が新たに必要になる。perturbation model でも、他のモデルと同じように、「ユーザと教師は少なくとも比較可能な程度に似ている」という仮定を置いているが、この仮定は直観的にも問題はない。ユーザと教師はある程度共通の背景知識を持っているはずで、そうでなければコミュニケーションなど成り立たないからである。

3.4.4 適用事例

前節で紹介されている日本語会話CAI システムに、上で述べたようなタスク/ユーザモデルを適用してみる。まず、タスクモデルは、学習者カード、文法カード、スクリプトカードに分散されて記述されている。学習者カードはシステムのある状態におけるユーザの大まかなタスクが書かれている。また文法カードは、その状態においてユーザの発話すべき文の文法的知識が書かれている。またスクリプトカードはタスクの細かい内容、アクションの手順などが書かれている。このようにタスクに関しては、これらの枠組でだいたい捉えることができる。

しかしこのままでは、特定のユーザ(学習者) に合わせたきめの細かい指導をおこなうことができないため、新たにユーザモデルに相当する個人カードを導入する。

個人カードは、たとえば次のようなものである。

個人カード	
カード ID	P01
個人名	佐藤一郎
音声モデル	男性 15
知識レベル	Novice
修了コース	SC008, SC009
習熟レベル	B, A

図 3.4 - 1 個人カードの例

ここで、知識レベルは、{Novice, Moderate-Expert, Expertg} などのようにその分野や項目に詳しいかどうかを自己申告させる。これによって、どの程度高度な内容を学習させるかどうかが決まる。また、習熟レベルは、修了したコース(タスク) がどの程度うまくいったかどうかを表す。A、B、C の順でうまくいったものとする。これは、コース修了時点で学習者が自己評価する。あるいは、システム側がユーザの発話がどの程度タスクモデルと整合したかどうかに合わせて自動的に評価するのでも構わない。ただし、この場合の評価には、音声認識のモデルがそのユーザの声にどれだけ合っていたかなどの要因も絡んでくるので、多少困難になるだろう。

このようなユーザモデルを用いて、各々のユーザに適した学習ストーリーを決める。そのために、たとえば次のようにスクリプトカードをより詳細にする必要がある。

スクリプトカード			
カード ID	SC010		
スクリプト名	電話をかけて食事の約束をする		
項目	内容	(条件:カード)	実行済
自己紹介	自分の名と会社名を言う	((Expert, SC008 ≥ B):S010a), ((Novice, SC008 < B):S010b), ...	done
相手指定	話したい相手の名を言う	(():S011)	done
開始挨拶	はじめの挨拶をする	((Expert, SC009 ≥ B):S030a), ((Novice, SC009 < B):S030b), ...	done
...	.....	.....	...

図3.4-2 改良スクリプトカードの例

ここで、<条件:カード>という部分は、ある条件を満たした時にあるカードに進むことを意味しており、たとえば、<(Expert, SC008 B):S010a>は、その学習者の知識レベルがExpertで、スクリプトカードSC008 の習熟度がB 以上だったならば、カードS010a を実行せよ、ということを表している。また、<():S011>は条件部が空なので、無条件でカードS011に進むことを意味している。

このように、ユーザモデルとタスクモデルを組み合わせて用いることによって、より柔軟な対話アプリケーションを構築することができる。

参考文献

1. [Brown 78] J. S. Brown and R. R. Burton. Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 2:155-192, 1978.
2. [Carr 77] B. Carr. WUSOR-II: a computer-aided instruction program with student modeling capabilities. AI Lab Memo 406, MIT, 1977.

3. [Kass 89] R. Kass. Student modelings in intelligent tutoring systems _ Implicationfor user modeling. In User Models in Dialog Systems, chapter 14, pages 386-410. Springer-Verlag, 1989.
4. [Schank 75] R. C. Schank. Conceptual Information Processing, volume 3 of Fundamental Studies in Computer Science. North-Holland, 1975.
5. [Schank 82] R. C. Schank. Dynamic Memory. Cambridge University Press, 1982.
6. [Schank 77] R. C. Schank and R. P. Abelson. Scripts, Plans, Goals and Understanding. Lawrence Erlbaum Assoc., 1977.

第４章　現状音声認識技術適用時の留意事項

第４章　現状音声認識技術適用時の留意事項

４．１　音声認識応用システム開発のガイドライン

４．１．１　はじめに

本節は、現在利用可能な音声認識･理解技術を、実際に応用（アプリケーション）システムに組み込む際に注意すべきノウハウ的留意事項について述べる。

音声認識･理解技術を応用システムに組み込む際には、音声認識･理解技術の現状について適切に理解する必要がある。音声認識･理解応用システムの分類については第２章に、音声認識･理解技術を組み込むための支援ツールについては第３章に、認識語彙を選択するための支援ツールについては第４章 4.2節に、音声認識･理解システムの評価については第５章に、標準化動向については第６章に詳しく述べられているので参照されたい。本節では、こうした点については簡単にふれるのみとし、主に、他の章で扱われていないノウハウ的な留意事項について述べる。

読者としては第１に、音声処理について経験はないが、音声を用いた何らかの応用 (アプリケーション) システムを開発しようと考えているシステム開発者を想定している。また音声処理について経験はないが、音声処理を統合したシステムについて研究しようと考えている研究者にも参考になるようにした。

さて、音声認識･理解技術に対する誤ったイメージの典型的なものはつぎのようなものであろう。

（１）音声認識･理解システムは万能であって、人間と同じような音声認識･理解能力がある。したがって人間に対するのと同じように話しかけてもちゃんと正しく認識･理解できる。

（２）カタログ上では95％の認識率となっているのに実際に使用してみると、ずっと認識誤りが多く、まったく使いものにならない。

これらのイメージはどちらも誤解である。

すなわち、現在利用可能な音声認識･理解技術を、実際に応用システムに組み込む際に注意すべきことは、音声認識･理解システムの仕様を適切に決める必要があるということである。音声認識･理解技術の発展により技術レベルはかなり高くなってきてはいるものの万能ではない。一方、実際に使用する際の認識性能は、音声認識･理解システム単体の性能だけでなく、その使用法や周辺を含めたアプリケーション全体の構成によって大きく左右される。

音声認識･理解システムの仕様を適切に決め、またいくつかのノウハウ的な留意事項に注意

することで、全体としての認識性能が大きく改善されるのである。

4.1.2 音声認識応用システムの開発において考慮すべき項目

音声認識応用システムの開発においては、表4.1-1に示したようなさまざまな項目について考慮し、その応用に適した仕様とする必要がある。

表4.1-1　音声認識応用システム開発における留意事項

---

(1) 音声認識の形態 1)~3)

1) 話者：特定話者／話者適応／不特定話者

2) 発声単位：孤立単語／連続単語／文節／文

3) 語彙：認識対象語数、単語の選択と組合せ

---

(2) タスクの設計における留意事項 4)~6)

4) 文法、文型の定義

5) 対話制御の設計

6) ユーザインタフェース

---

(3) その他の留意事項 7)~9)

7) 収録系

8) システム評価法（テスト法）

9) ハード構成 (音声認識ボード／音声認識装置／音声認識 LSI)

（1）音声認識の形態

音声認識応用システムの仕様を決める際にはじめに考慮することは、音声認識の形態である。一般に、音声認識の形態に制限を加えず自由なものとすればそれだけ音声認識は難しくなり、認識性能が低下する。逆に、音声認識の形態に制限を加えれば、認識性能は向上する。したがって、アプリケーションの性格から決まってしまう項目についてはやむを得ないが、そうでない項目についてはある程度制限した方がよい。

認識性能に影響する形態の主な項目は、話者、発声単位、語彙の3つである。これらの項目について以下に簡単に述べる。音声認識の形態と性能についての詳細は、第2.4節を参照のこ

と。

1) 話者：特定話者／話者適応／不特定話者

対象とする話者によって、特定話者型（複数特定話者型を含む）、話者適応型、不特定話者型の３つに分類できる。「特定話者型」の音声認識システムは、使用する人（話者）にあらかじめ話してもらって、その声を登録しておく方式である。「不特定話者型」の音声認識システムは、使用する話者の発声を事前に登録しない方式である。「話者適応型」の音声認識システムは、使用する話者のすべての言葉（単語）を登録するのでなく、一部の言葉だけを話してその話者に適応させるような方式である。

一般に認識性能としては、認識対象語数などの他の条件が同じであれば、

不特定話者＜話者適応＜特定話者

の順に認識率が高くなる。したがって、誰が話すかをあらかじめ特定できないような応用では「不特定話者型」とせざるを得ないが、使用する話者が特定できる場合には、できるだけその話者の発声を事前に登録するか、あるいはその話者に適応させるようにした方がよい。

特定話者型、話者適応型、不特定話者型のどの場合にも、認識しやすい人と認識しにくい人があり、使用する話者によって認識性能はかなり変化する。使用する人のうち、だいたい８～９割くらいの人の認識率は高いが、残りの人の認識率はかなり低下するという傾向がある。

また、話者が特定できるような応用では、話者の方が認識システムに慣れるということもあるので、実質的にはさらに認識率が高くなる。たとえば、音声認識システムを導入した当初は、話者も緊張して認識率があまり高くないように見えるが、認識システムに慣れるにしたがって認識率はかなりの程度向上する。

2) 発声単位：孤立単語／連続単語／文節／文

音声認識システムの発声単位には、孤立単語、連続単語、文節、文などがある。孤立単語発声や文節発声では、単語や文節ごとに区切って発声するという制約があるが、文単位の発声を対象とする連続音声認識に比べて、単語認識率での認識性能にはかなりの差がある。

単語や文節ごとに区切って発声することは、話者には負担となるが、現在実用化されている音声認識システムの多くのものは、孤立単語発声を基本としている。たとえば、項目を選択するような応用であれば、単語で指定するように設定する。

文単位の発声を対象とする連続音声認識の場合にも、使用する文型を適切に制限しておけば、最終的な認識率を高めることができる。この場合には、使用する単語数よりも対象とする文型

の複雑さで認識性能が決まる。

音声認識システムに対して、人によっては「えーっと」、「えー」、「あのー」などの言葉（いわゆる不要語）を発声してから認識対象とする言葉を発声することがある。現在実用化されている音声認識システムの多くのものは、こうした不要語をうまく扱えないが、認識システムに話者が慣れるにしたがって、かなりの程度までは不要語を発声しないようにすることができる。またこうした不要語をどうしても発声しがちな話者の場合などには、不要語も含めて登録しておくことも１つの解決策である。

3) 語彙：認識対象語数、単語の組合せ

認識システムが対象とする単語の数が多い程認識率は低下し、少ない程認識率は高くなるので、認識対象語数はアプリケーションからの制約と認識性能とのバランスで決まる。現在の典型的な音声認識システムの形態を例示すれば、つぎのようになる。

特定話者 or 話者適応：孤立単語発声：認識対象語数 1000 語以内
特定話者 or 話者適応：連続単語発声：数字＋コマンド程度
特定話者 or 話者適応：文節 or 文発声：認識対象語数 100 語程度
不特定話者：孤立単語発声：数字＋コマンド程度

このような形態でどの程度の認識率が得られるかは、使用する語彙、文型についての制約、収録系、個人差などさまざまな要因があって一概には決められない。あえて言えば、このような形態で 95%前後の単語認識率というのが典型的な数値である。

同じ認識対象語数であっても、使用する語彙が特定の語彙に固定されている場合と任意の語彙を対象とする場合では認識率にかなりの差がある。たとえば、数字の認識は、読み方にいろいろあることやあまりていねいに発声されないことが多いなどの理由から、みかけの語数から想像されるよりずっと難しいタスクである。しかし実用的なニーズが高く、学習用の音声データもかなり蓄積され、また盛んに研究されているので、かなり高い認識性能が得られている。電話回線を通した不特定話者の単語音声認識システムが実用化されていることにも、数字と機能語という特定の語彙に限定していることが大きく寄与している。

応用システム側から使用する語彙が決まるような場合には、こうしたことが期待できない。単語単位で話者の音声を登録する特定話者型の音声認識システムでは、使用する語彙内の単語をすべて発声してもらい、それを登録することになる。また認識対象とする語彙を、音韻（音素）列で指定したり、あるいはシステム側で用意した辞書から選択したりするものもあるが、特定の語彙を対象とする場合に比べると認識性能はある程度低下する。

認識対象とする語彙を決めることは、意外と難しいことである。対象とする応用を考えて最初にある程度の語彙をリストアップし、さらに模擬的なシステムでの対話例などを収集したりするのがよい。多くの場合、それでも十分ではなく、実際に使用しながら語彙を追加したり、変更したりする必要が起きることを想定しておく方がよい。

また、単語の組合せを考慮して、認識誤りしやすい語は除くといった認識対象語彙の選択については、第 4.2節にさらに詳しく述べられている。

（2）タスクの設計における留意事項

4) 文法、文型の定義

連続音声を対象とする場合には、文法を記述する必要がある。音声認識システムで使用される文法は、通常の場合、単語または文法的カテゴリーのネットワーク表現で表される。現在の連続音声認識システムでは認識性能の点で、大語彙のものは難しい。小語彙であれば、文型をネットワーク表現などで表すことは簡単で有効な方法である。文脈自由文法で記述される場合もあるが、応用システムのタスクや領域に適した文法を書くことがかなり難しいので、その点に注意が必要である。

日本語の一般的な文法を書いたとしても、音声認識システムのための言語的制約としては弱いので、実際的には、タスクに適したもっと制約の強い文法が必要である。

応用システムのタスクや領域でのシミュレーションを行って、そこに出現した文型を収集して、それを音声認識システムで使用する文法とするという程度の手間をかけた方がよいものが得られる。

5) 対話制御の設計

使用者の発話とシステム応答が交互に繰り返し行われるような応用システムでは、そうした対話の流れやシステム応答を設計する必要がある。はじめに応用システムのタスクにそったシナリオを設定して、それに応じて対話の流れやシステム応答を決める。この段階での検討が重要である。

対話の流れの設計は、システム主導であるか、ユーザ主導であるかによっても異なるが、システム主導であっても、対話の途中でのキャンセルなどの最低限のコントロールはユーザが行えるようにした方が、より柔軟なシステムとなる。

認識した結果として、リジェクト（棄却）を許すことは、音声認識システム全体の信頼性を高める効果がある。この場合にもユーザにリジェクトであることを知らせた方がよい。

応答メッセージの生成においては、質問の形式に一貫性をもたせることや、ユーザ発話を正しく認識していることを示すために認識結果の一部を応用メッセージに入れるとともに全体と

しては冗長にならないように注意することなどの工夫が必要である。音声応答の場合には冗長にならないように注意することが、特に重要である。

詳細は、第3章音声対話アプリケーション開発支援ツールの節を参照のこと。

6) ユーザインタフェース

音声認識応用システムのユーザインタフェースによっても認識性能はかなり左右される。

第1に発声のタイミングの制御が重要である。たとえば、開始を指示する方法には、ボタン、マウス、ピッという合図音などがあるが、いずれもある程度の慣れが必要である。システムの名前を呼びかけることで、発話の開始を指示する方式もある。またキーワード発声などのように、ユーザの発話をつねに受け付ける方式もあるが、ユーザ発話の音声区間を切り出す処理や認識対象となる単語だけを取り出す処理においてある程度の切り出し誤りが起こるので、開始を直接指示する方式に比べてそれだけ全体の認識性能が低下することになる。

第2に入力音声のパワーのコントロールが以外と重要である。ユーザの声の大きさとマイクとの距離の両方に注意する必要がある。ある程度慣れも必要であるが、入力音声のパワーを表示することなどが効果的である。

第3として、マウス、グラフィックス、キーボード、タッチパネルなどの他の入出力デバイスと統合することで、全体としての使いやすさと認識性能が向上するということがあげられる。たとえば、認識結果の第1位の候補のみを出力するのでなく、第N位まで出力してその中から選択することで、全体としての認識性能を大きく向上させることができる。

（3）その他の留意事項

7) 収録系

応用システムとして、電話回線を使用するかどうかで、利用可能な音声認識システムも異なる。使用するマイクは、接話型／卓上型、いずれも可能であるが、認識モデルの学習時と認識時で異なるマイクを使用し、かつその対策をとくにとらない場合には、予想外の認識性能の低下が見られる場合があるので注意を要する。フィルターについても同様の注意を払う必要がある。

周辺雑音の問題に対して、別のマイクを用意してそれから得られた音を利用してノイズなどを除去することは、特に雑音レベルの高いような環境での認識性能の向上にとって有効である。また音声認識システムを使用する環境での周辺雑音を、実際に使用する収録系（マイクとフィルター）であらかじめ収録して、認識モデルをその使用環境に適応させたり、音響的な特性を補償したりすることが有効である。

標本化周波数は、多くの場合、音声認識システム側で決まっているが、電話などで使用され

ている標本化周波数 8kHz では、それ以上の標本化周波数に比べて認識が難しい。また計算機側に内蔵されているマイクやフィルターを使用する場合には、これらの項目については、選択の余地がない。

8) システム評価法（テスト法）

音声認識システムのカタログ上の性能は目安であって、実際に応用システムが使用される環境で得られる性能は必ずしもカタログ性能と同じではない。また学会などで発表されている研究所内などでの実験結果と、実際のフィールドでのテスト結果にはかなりの差があるので、学会で発表されている数値はその点について注意する必要がある。

カタログ上の性能を求めるのに使用された音声データや評価法が同じでない限り、いくつかの音声認識システムの性能を見かけのカタログ性能で比較することにはあまり意味がない。共通的に利用可能な音声データには、音響学会、電子協、ATR から公開されているものなどがある。音声データベースの現状と動向については、昨年度の報告書にさらに詳しく述べられている。

また同一の音声データと評価法で比較した場合にも、統計的に意味のある差とするためには、テストサンプル数を十分に多くとる必要がある。認識率のわずかな差であれば、それは差がないと考えるべきである。

応用システムを開発する課程では、音声認識システムの性能を評価し、その結果をシステム設計に反映させることは重要であるが、認識性能だけでなくユーザインタフェースまで含めた「ユーザの満足度」という観点から評価する必要がある。詳細は、第 5 章音声認識・理解システムの評価の章を参照のこと。

9) ハード構成 (音声認識ボード／音声認識装置／音声認識 LSI)

現在入手可能な音声認識関連製品をハードウェアの構成で分類すると、音声認識ボード、音声認識装置、音声認識 LSIの 3 つに分けられる。入手可能な音声認識関連製品の一覧を資料編に示す。

ここにあげられた音声認識関連製品は 82 種類あるが、そのうち、音声認識ボードは 31 種類、音声認識装置は 36 種類、音声認識 LSIは 15 種類となっている。音声認識ボード／音声認識装置／音声認識 LSI のいずれを採用するかは、音声入力を接続するホスト側の機種 (PC / WS) とコストなどのかねあいによって決まる。

近年の PC/ WS では、計算機側にオーディオ入力が標準装備される傾向にあり、また、計算機本体の演算速度と利用可能な記憶容量が増える傾向にあるので、ソフトウェアのみの音声認識システムも可能になりつつあるが、まだ製品レベルにはなっていない。

この場合、音声認識システムに必要な演算速度と記憶容量は、認識方式によってもかなり変化するが、最近普及している WS 程度の性能 (たとえば 50MIPS, 16 MB)があれば、前記に例示した典型的な音声認識システムはソフトウェアのみで実現可能である。

4.1.3 音声合成システムの開発において考慮すべき項目

この項では、音声合成システムの開発において考慮すべき項目について簡単に述べる。

音声合成関連製品一覧を資料編に示す。音声合成の方式としては、大きく分けると、録音再生（分析合成を含む）／規則合成の２種類がある。

録音再生は、あらかじめ録音しておいた音声を、指令に応じて再生するもので、情報圧縮の方式には、ADM (Adaptive Delta Modulation), ADPCM (AdaptiveDifferential PCM), CELP などの符号化法として知られるものや、波形編集、分析合成方式などから出てきたものがある。分析合成のパラメータとしては、ホルマント、PARCOR (PARtialauto CORrelation)、LSP (Line Spectrum Pair)、メルケプストラムなどが用いられている。

録音再生型の音声合成には多くの種類の専用 LSI が発表されており、さまざまな家電製品、玩具、エレベータなどで使用されている。この方式では、合成音の音質はかなり高いが、あらかじめ録音しておいた音声を再生するために、合成できる言葉の種類が記憶容量によって限定される。

一般には通信伝達が主目的の符号化を行う方式に比べて、波形編集や分析合成を行う方式のものの方が情報圧縮の程度が大きい。利用可能な記憶容量が同じであれば、一般的な傾向として、情報圧縮の程度が大きい程、合成できる言葉の種類はそれだけ多いが、音質の点ではそれだけ低下する。

規則合成型の音声合成は、テキストから任意の言葉を合成するものである。製品としては、PC用のボードや RS232C インタフェースを介して接続する専用装置などがある。音質は、録音再生型の音声合成に比べれば低下するが、おおむね良好で、合成された言葉をだいたい了解できる程度の音質はある。

入力の形態は、漢字かな混じり文のものと、韻律記号付きかな文字列のものがある。漢字かな混じり文を入力とするものは、システム側で言語処理を行って韻律を制御する。使用できる辞書の大きさは製品によっても異なるが、２万語から10万語程度である。発声速度は10段階程度で調節可能であり、声の高さを変えたりできるものもある。

音声合成システムの評価は、基本的には難しい問題であり、多くの課題がある。合成音声の了解性と自然性とともに、実際に使用される環境を想定した総合的な評価が重要である。最近、音声合成システムの評価法について、精力的に研究されており、海外においても活発に研究されている。

最後に、音声合成のインタフェースは現在のところ各社ばらばらであり、早急な共通化が望まれる。

4.2 使用語彙の選択支援ツール

4.2.1 目 的

音声認識·理解技術を応用するシステムを開発している者に、音声入力用に計画している用語の中から、認識誤りの多い言葉などを予測、警告，改善案提案などをする手段（ツール）を提供するのが目的である。

この使用語彙選択支援ツールは、使用者が特定の数名のメンバに限定される応用では、用語の制限が可能なため、大きな効果を発揮するものと期待される。一方、不特定話者方式では、発声許容語彙を制限することは一般に困難なため、誤りが少なくなるように処理の流れを見直す手段として利用できる。言い替えれば、誤りの起きやすい用語が発声される確率が低くなるようにアプリケーションの処理手順を変更したり、発声者への適切な指示、誘導を行う処理を付加するなどの対策を打つきっかけを与えることができる。

なお、本ツールは、音声インタフェース開発支援ツールの補助機能としても、組み込まれていると非常に便利である。

4.2.2 使用語彙選択支援ツールの機能

音声認識·理解機能を有するアプリケーションシステムで用いる語彙を選定する場合、提供されることが望ましい機能には、次ぎのようなものが考えられる。

警告機能　与えられた認識用単語の中から、誤りの起こりやすい組み合せを検出し警告する［江袋 87］。

選択機能　認識用単語を類義語の群ごとにまとめて与え、全体として誤りの少ない組み合せを各群の中から選択、提案する［矢島 90］。

提案機能　類義語辞書を用意し、与えられた認識用単語ごとに類義語を検索し、全体として誤りの少ない組み合せとなるよう、各類義語の中から自動選択、提案する［石垣 86］。

後の例ほど機能は高度となっている。以下に各引用文献の提案などを参考に、その構成例を説明する。

（1）警告機能

音素系列などによる単語の表記法と、この表記法を用いて表記された任意の1組の単語の間の類似度（あるいは、距離、相違度など）の予測手法をなんらかの方法で求めておく。これを用いて標準単語として登録されている単語の間で誤認識の起こりやすい組み合せを予測し、登録単語を変更するよう警告する。

単語の表記法は、多くの人が利用できるように、発音記号などより、ローマ字表記や仮名表記が望ましいと思われる。本ツールでの内部表記は、無声化や脱落、異音、音便などの併記、音韻間のわたり部での音韻湧き出しやポーズの付加、などの情報を付け加えた発音記号的なもののネットワーク形態が考えられよう。

（2）選択機能

警告機能と最も異なる点は、音声認識･理解に用いる用語を、用語ごとに異なった読み（たとえば、7のナナとシチ、など）や、同義語（計算機とコンピュータなど）をセットとしてアプリケーションシステムの設計者がツールに与える点である。

用語の誤認識の可能性の評価は警告機能と同様に行う。その結果をもとに、相互の誤認識ができるだけ小さくなるように、各用語セットごとに1つないし幾つかの用語候補の選択を行う。

（3）提案機能

この機能では、更に進んで、ツール自体が音声認識･理解に用いる用語として、用語ごとに異なった読みや同義語を、自身の内部に用意されている辞書から捜し、相互の誤認識ができるだけ小さくなるように、各用語セットごとに1つないし幾つかの用語候補を選択し、提案を行う。

4.2.3 開発を要する技術

（1）文字列からの音声誤認識の推定法

音韻間の類似度や単語間の類似度（あるいは距離、相違度）の最適な推定方式は、使用する音声認識･理解装置の方式に合わせて変るべきものである。しかし、簡便性や可搬性を考えると、多くの装置に対し、できるだけ性能上のずれが生じないように配慮された共通方式にしておいた方が実際的と考えられる。一般的な音韻学的知識に基づいた方式などが考えられよう。

このような観点から見た場合、音韻間の類似度は、ポーズや全ての異音を含めた音韻間のコンフュージョン・マトリックスで与えるなどの方法が考えられる。

単語間の類似度は、記号ネットワークを対象とした両端フリーの時間軸非線型マッチングを利用することが1つの方向であろう。異音など複数の候補のいずれかが生ずる可能性のある部分では、それらのうち最も類似度の高いものの組み合せを、また湧きだしの可能性のる部分で

は、湧きだしの有り無しのうちから類似度の高い方を選択して評価する。これにより、最悪状態を近似評価することが可能となる。

（2）最適組み合わせ選択手法

相互の誤認識が最小になるように、各用語セットごとに1つ用語候補を選択する手法の開発が必要である。以下に一試案を示す。

同一セット内の組み合せを除く全ての候補単語の間の類似度マトリックスを求めておき、類似度の最も高い組み合せを取り出す。その組み合せの2単語の各々について、2番目に類似度の値の高いものを取り出し、両者を比較して値の高い方を候補単語から外す（2番目の値が同じ時は3番目で比較、以下同様）。同一セット内の単語の数が1つ（あるいは、あらかじめ定めた数）になったら、そのセットに属する単語は外す対象からは除外し残す。この操作を繰り返し、全てのセットについて、各々予定していた単語数に絞られたら、その時点で終了する。

この処理は事前に行うものであるから、処理速度は大きな問題ではないが、初めに全ての候補単語の間の類似度を値の順にソートしておくなどにより処理規模を押えることも可能であろう。

（3）類義語辞書開発

提案機能では、用語毎の異なった読み方（簡略化した読み方などを含む）や類義語辞書（シソーラス）の開発が必要である。しかし、一般性のある辞書の開発は非常にコストがかかり、記憶容量も大きくなる。なんらかの自然言語処理用に開発されるものを流用すべきであろう。

参考文献

1. [江袋 87] 江袋　林蔵、”相異度法による単語の相異と音声認識性能の解析”,電子情報通信学会論文誌A、Vol.J70-A、No.11、 pp.1533-1544、1987.
2. [石垣 86] 特許公開公報　昭６１－２６０９６”音声認識単語の事前評価システム”
3. [矢島 90] 特許公開公報　平０２－３０４４９５　”音声認識単語セットの評価システム”

第５章　音声認識･理解システムの評価

第５章　音声認識･理解システムの評価

音声認識･理解システムに関して精力的に研究が行われており、今後様々な音声認識･理解システムが実用化されると考えられる。実用化を行う場合、音声認識･理解システムの評価を行い実用化レベルであるかどうかを判断する必要がある。判断を行う評価尺度として一般的には認識率が用いられている。しかし、認識率が良くてもシステム全体の使い勝手が悪いとユーザに受入れられない。また、認識率が悪くても認識結果の上位ｎ個を表示させ、ユーザにタッチパネルなどで選択させれば、実用に耐えられる場合もある。したがって､ユーザに音声認識･理解システムを満足して使用してもらうためには、単に認識性能だけではなく、ユーザインタフェースまで含めた「ユーザの満足度」を評価の基準とする必要がある。「ユーザの満足度」は、厳密に考えれば主観的な評価になるが、ある程度客観的に評価できるように「ユーザの満足度」の評価項目についての提案を行う。

５.１　「ユーザの満足度」について

音声認識･理解システムを評価するための評価項目について提案を行う。ユーザは実際に使用するエンドユーザと、システムを導入するユーザの２つがある。エンドユーザは”使いやすさ”で満足度を判断するが、システムの導入者は音声認識･理解システムを導入することによって”仕事が経営的にどの程度改善されるか”で満足度を判断する。

表５.１-１にエンドユーザの意見をまとめる場合の評価項目、表５.１-２にシステム導入者に対する評価項目を示す。各項目に関してどの程度重視するかは場合によって異なると考えられるので、ここでは触れないことにする。以下に各項目についての説明を行う。

（１）音声理解システムの目標達成に関して

以下のような項目を評価し、従来のシステムと比較を行う。

(a)目標達成までの時間

音声理解システムとＱ＆Ａを繰り返し、目標が達成できるまでの時間を測定し、平均時間を求める。

(b)目標達成までのインタラクション数

システムと繰り返したＱ＆Ａの回数をカウントし、色々な場合の平均回数を求める。

(c)目標達成率

使用者の中で最終目標まで到達できた人の割合。

(d)目標に対してどの程度情報が得られたか

目標が達成できなかった人が得た情報の割合。目標の情報が１つの場合は、０％か100％になる。

表５．１-１　ユーザに対する満足度評価項目（まとめ）

(1) 目標達成に関して

(a) 目標達成までの時間	平均	秒
(b) 目標達成までのｲﾝﾀﾗｸｼｮﾝ数	平均	回
(c) 目標達成率	平均	%
(d) 目標に対してどの程度情報が得られたか	平均	%

(2) 発声に関して

(a) 発声のタイミング	自然		不自然
(b) 不要語、言い淀みの発声	可		不可
(c) 発声の言葉に対する制限	無		有

(3) 認識性能

(a) 認識性能	満足		不満足
(b) 言葉による認識率のばらつき	少ない		多い
(c) 話者による認識率のばらつき（最高と最低認識率の差）			%

(4) 使いやすさ

(a) 使いやすさ	良		悪
(b) マニュアル（説明）なしで使用できる	可		不可
(c) 状態遷移 (Q&Aなど) の自然さ	自然		不自然
(d) 常にどの状態にいるかがわかりやすい	容易		難
(e) 誤認識した場合の修正	容易		難

(5) 感想

(a) また、使いたいと思う人の割合　　%

(b) 感想

表5.1-2　システム導入者に関する項目（経営的に）

(6) システム導入者に関する項目（経営的に）		
(a) 人の削減（人件費の削減）	大	小
(b) 仕事のコストパフォーマンスの向上	大	小
(c) 仕事の信頼性の向上	大	小
(d) 会社の宣伝効果（イメージアップ）	大	小
(7) 感想		

（2）発声に関して

(a)発声のタイミング

発声のタイミングは自然であることが望ましい。

・自　然：いつでもユーザの発声を認める。

・不自然：発声のタイミングを制限している場合。例えば、認識開始の合図音（ピッなど）を鳴らす。普段から使用している人に関しては問題とならないが、はじめて使う人に対してはかなりの精神的負担になる。

(b)不要語、言い淀みの発声

使いやすくするためには、発声の前後の不要語（「あー」、「えー」など）や言い淀みなどの発声を認める必要がある。

・可　：不要語を認める。

・不可：不要語を認めない。

(c)発声の言葉に関する制限

・無：発声者が発声する色々な言い回しに関しても認識できる。

①読み方

数字の「7」の発声を「なな」または「しち」のどちらの発声でも認識できる。

②略称

「新横浜」と発声したい場合、略称である「しんよこ」と発声しても認識できる。

③語順の入れ替え

文発声の場合、色々な語順の入れ替えにも対応できる。

・有：決められた言い回ししか認識できない。

（3）認識性能に関して

(a)認識性能

・満　足：実環境騒音が色々変化（ノイズの種類、S／N比など）しても高い認識率が得られる。

・不満足：環境騒音によっては、認識率が低下する場合がある。

音声認識の技術的課題である音声認識の性能に関しては、実際の使用時の認識性能が重要である。研究者が用いている評価データに対しては高い認識率が得られても実際に使用すると低下する場合がある。したがって、できるだけ色々なデータ（実環境をできるだけ忠実に再現したデータ）により評価し実環境に耐えられるように音声認識法を開発する必要がある。音声認識・理解システムが実際に使用される環境として、どのような条件を想定し評価すればよいかを慎重に検討する。この現実の環境をどれだけ正確に予測し、それに対処できるように認識法を改良するかで認識性能は決定される。

たとえば、事務所環境ではＯＡ機器や人の声など、自動車内ではエンジン音、風切り音などがあり低域にエネルギーが集中した場合や高域にエネルギーが集中した場合など色々な場合が考えられる。また、環境ノイズのＳ／Ｎ比、時間的変化になども重要な検討項目である。

(b)言葉による認識率のばらつき

・少ない：どの言葉（単語）に対しても一定の認識率が得られる。

・多　い：言葉によって認識率の変化が多い場合。

認識率の悪い言葉（単語）があれば何らかの対処をする必要がある。特に、単語数が多くなると類似単語が増えてくるので言葉による認識率のばらつきが大きくなる。

(c)話者による認識率のばらつき（最高と最低認識率の差）

平均の認識率より、認識率の悪い話者（最低認識率）がどの程度かが重要である。ここ

では、話者別認識率で最高認識率と最低認識率との差を評価尺度とする。当然のことながら、差が小さいほど良い。

また、一般的には発声話者として成人男女を想定して評価を行っているがアプリケーションによっては、子供から老人まで対象となることが考えられる。この子供や老人を対象とする場合は、かなり声の性質が異なるので検討を行う必要がある。

（4）使いやすさに関して

(a)使いやすさ

音声認識理解システムは使いやすくなければいけない。

・良：実際に使用した人の感想で「使いやすい」と思った人の割合が多かった場合

・悪：実際に使用した人の感想で「使いにくい」と思った人の割合が多かった場合

使いやすくするためには、研究者がユーザと一体になってシステム全体のユーザインタフェースを開発する必要がある。たとえば、第1位の認識率が悪い場合でも、上位n個の結果を表示してユーザにタッチパネルなどで選択させることによって使いやすくすることができる。このように使いやすいインタフェースをユーザと一緒に構築することが重要である。

(b)マニュアル（説明）なしで使用できる

音声認識は人間にとって最も自然なヒューマンインタフェースの1つであるのでマニュアルや説明なしで自然に使用できる必要がある。

・可　：マニュアル（説明）なしで使用できる。

・不可：マニュアル（説明）がないと使用できない。

(c)状態遷移（Q＆Aなど）の自然さ

音声理解システムにおいて、Q＆Aの遷移が自然であるかどうかが重要である。

・自　然：使った人に感想を聞いて状態遷移が自然であると思った人の割合が多い場合。

・不自然：使った人に感想を聞いて状態遷移が不自然であると思った人の割合が多い場合
特に、同じ様なQ＆Aを何回も繰返し聞いたり、思っていることを一度に発声できないようなシステムは不自然である。

(d)状態遷移がわかりやすい

・容易：目的達成までのQ＆Aの遷移に関して、自然な状況推移でいつでもユーザが現在

どの状態にいるか分りやすい。また、状態間の遷移が自由に移動できるようになっている。

・難　：どの状態にいるかが分かりにくい。

(e)誤認識した場合の修正

・容易：誤認識しても、必要最小限のやりとりで容易に修正できる。

・難　：誤認識すると修正が複雑である。

（5）使用者の感想

(a)また使いたいと思う人の割合

使った人に感想を聞いて、これからも音声認識·理解システムを使用したいと思った人の割合。

(b)感想

「ユーザの満足度」はやはり実際に使用した人の感想が最も重要である。感想をまとめて今後の参考意見とする。システム開発者は、できるだけユーザの感想を聞くようにして、改良を行っていく。

（6）システム導入者に関する項目（経営的に）

音声認識システムが低価格で個人ユースのものは、使用者も導入者（購入者）も同一人物あるが，音声認識·理解システムとして企業などが購入して使用する場合は、導入者に関する満足度は実際のユーザとは異なったものになる。導入を検討する人に関しては、使い勝手よりも導入することによって経営的にどのようなメリットがあるかが、満足度の重要な項目となる。

(a)人の削減（人件費の削減）

システムを導入することにより、従来システムより人（人件費）の削減がどの程度できるかを評価尺度とする。

(b)仕事のコストパフォーマンスの向上

仕事の処理速度が上がりコストパフォーマンスが向上したか。

(c)仕事の信頼性の向上

音声を使うことにより仕事の信頼性が向上したかを評価尺度とする。たとえば、「データの入力間違いが少なくなる」など。

(d)会社の宣伝効果（イメージアップ）

先端技術の音声認識·理解システムを導入することによって、企業イメージがアップする場合がある。ここでは、イメージアップにどの程度貢献したかを評価尺度とする。イメージアップによって企業の採用活動に貢献する場合もある。

5.2 まとめ

以上述べた項目に対して検討がどの程度行われているかによって、客観的に「ユーザの満足度」の評価を行うことができる。音声認識·理解システムを実用化する場合には、認識性能だけではなくユーザインタフェースでかなり使いやすさが決定されるので、開発者は常に単純に認識性能だけで評価を行うのではなく「ユーザの満足度」でシステムの評価を行うように心掛ける必要がある。

第６章　音声処理技術の標準化

第6章　音声処理技術の標準化

6.1　音声処理技術標準化の背景

コンピュータの世界は、長い間、数値・記号データを処理の対象としてきた。しかし、処理速度・メモリ容量が飛躍的に向上したのに伴い、近年、処理対象を他のメディアに拡大する試みが進んでいる。人間同士の対話が様々なメディアを使用していることを考えると、コンピュータとの対話チャンネルがマルチメディア化することは、自然な流れである。他方、デジタル通信の世界においてもネットワークの高速・大容量化に伴って、音声・文字・画像のマルチメディアを同時に伝送する、デジタルネットワーク統合サービス（ISDN）の整備が進められている。ここでは処理の対象が音声から、文字・画像を含むマルチメディアに拡大されるに伴い、ATM（非同期伝送モード）などの新しい技術が開発されている。表6.1-1は、これらの要素技術と、対応する音声メディア関連技術を示している。

表6.1-1　マルチメディア処理技術と音声

マルチメディア処理技術		音声メディア関連技術
A．マルチメディアの蓄積技術	：	音声ファイリング
B．マルチメディアの伝送技術	：	音声通信
C．マルチメディア間の相互変換技術	：	文－音声変換，音声－文変換
D．マルチメディア統合技術	：	文・音声・画像間の同期など
E．マルチメディア利用技術	：	音声入出力とペン入力，グラフィックス出力などを統合したＨＩ

表の中で、A.とB.は主に符号化の技術（音声では圧縮・伸長技術）、C.は音声合成（特に規則合成）と音声認識技術を指す。また、D.は画像と音声間の同期など、異なるメディアを統合する場合に必要となる技術、E.はヒューマンインタフェース（HI）の技術、ことにマルチモーダル対話技術と呼ばれるものである。

処理の対象がマルチメディアへと拡大された結果、データフォーマットやプロトコルに関する規約を決める必要が生じている。ここでは、マルチメディア処理技術や音声処理技術の標準化の動向と、その将来について解説する。

6.2　マルチメデイア標準化の動向

マルチメディア標準化の中では、マルチメディアの蓄積・伝送に関わる符号化技術に関するものが最も進んでいる。表6.2-1にこれまでに決められた符号化に関する主な国際標準をまとめた。

表6.2-1　マルチメディア符号化技術に関する国際標準

メディア	名　称	内　容	標準化団体
音　声	G.711	log−PCM(64kb/s)	CCITT
	G.726	ADPCM(32kb/s)	CCITT (1988年)
	G.727	エムベデッドADPCM ATMセル転送用	CCITT
	G.728	LD−CELP(16kb/s)	同　上 (1992年)
	検討中	8kb/s版	同　上
オーディオ	MPEG.Audio	ステレオ	ISO
	G.722	SB−ADPCM 帯域7kHz, 64kb/s	CCITT
静止画	JPEG	カラー, ADCT	ISO, CCITT (1991年ドラフト)
動　画	MPEG	汎　用	ISO
	MPEG1:	～1.5Mb/s 動き補償+DCT	(1991年勧告)
	MPEG2:	CDROM, 通信・放送 ～10Mb/s;100Mb/s (HDTV)	(1994年勧告予定)
	MPEG4:	8kb/s以下, 移動無線	(1998年勧告予定)
	H.261	DCT テレビ電話(64kb/s) テレビ会議(384kb/s) テレビ会議(1.5Mb/s)	CCITT (1990年)
	H.26X	ATM対応(B−ISDN) (MPEG2との互換性課題) 10Mb/s	CCITT (1994年勧告予定)
文　書	ODA	文書一般	ISO, CCITT
	MHEG	マルチメディア文書	ISO

CCITT: Comite Consulatif International Telegraphique et Telephonique

ISO: International Organization for Standardization

ADPCM: Adaptive Differential PCM

LD-CELP: Low Delay Code Excited Linear Predictive Cording

SB-ADPCM: Sub Band ADPCM

MPEG: Moving Picture Experts Group

JPEG: Joint Photographic Experts Group

ADCT: Adaptive Discrete Cosine Transform

ATM: Asynchronous Transfer Mode

B-ISDN: Broadband-ISDN (Integrated Services Digital Network)

ODA: Open Document Architecture

MHEG: Multimedia and Hypermedia Experts Group

一方、マルチメディア統合技術および利用技術は、パーソナルコンピュータ（PC）やワークステーション（WS）の世界で、デファクトスタンダード（DFS）の形で標準化に向けての競争が繰り広げられている。マルチメディアPCの主なOSは、マイクロソフト社のWindows MMEとアップル社のQuick Timeである。これらのOS環境により、アプリケーション構築者はマルチメディアの応用ソフトを作成する場合、個々のメディア制御から解放されることになった。ただし、MME版、Quick Time版と複数のバージョンがあるため、セッション制御の部分は別に作成する必要がある。これに対して、Kaleida社はこの部分（セッション制御）のScriptを提供して、プラットホームに依存しない標準化、オープン化を目指している[1]。

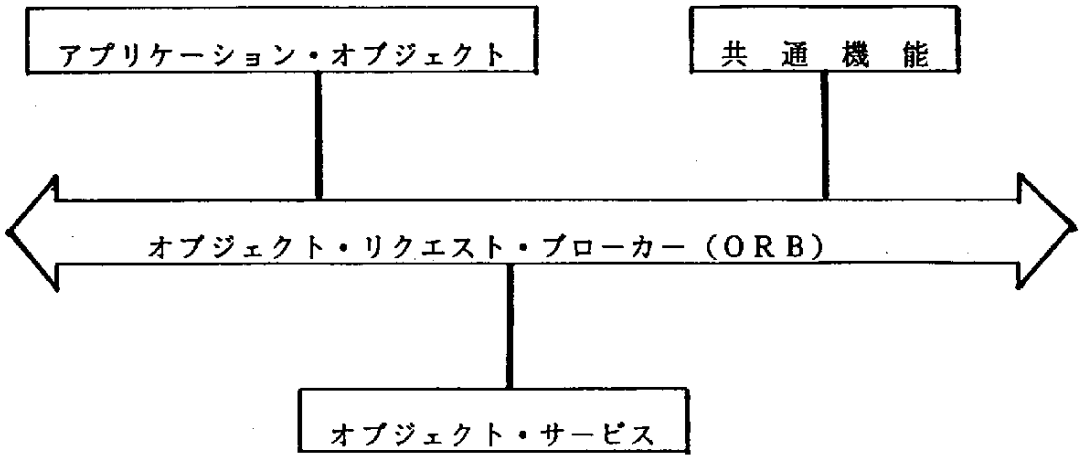


図6.2-1　OMG/ORB (Object Management Architecture)

一方、WSの世界ではOMG（Object Management Group）を中心にマルチメディアを取り扱う枠組みを分散オブジェクト指向技術によって構成しようとしている。OMGが定めるOMA（Object Management Architecture）は、図6.2-1に示すようにオブジェクト・サービス、アプリケーション・オブジェクト、共通機能、およびインタフェースの規定であるORB（Object Request Broker）から成る。最終目標はオブジェクト指向技術を分散コンピューティング環境に導入することにあり、これを実現するための規約は、CORB（Common ORB）としてオープンになっている[2]。

マルチメディアの標準化を広く推進しているIMA（Interactive Multimedia Association）は1987年に設立された業界団体で、現在、PCやWSメーカーを中心に約300社が参加している。IMAの目的はマルチメディア分野でのポータビリティを妨げる障壁を取り除くことで、次の5つのFOG（Focus Group）が構成されている。

A. Digital Audio
B. Data Exchange
C. System Services
D. Scripting
E. Text and Graphics

各FOGはRFT（Request For Technology）を定めて、それにしたがったアプリケーション／プラットホームの設計と実装の提出を広く業界に求めている。RFTでは、「～を実現する機能」といった形式で必要な条件を定め、API（Application Program Interface）等については規定しない。これはRFTがプラットホームとして唯一の環境を求めるものではなく、標準を土台とした様々なプラットホームから、ユーザーが必要なものを選択できる市場を構築することを目的としているからである。上記の中でC.のSystem Servicesは、WSの有力メンバーが入っており、前述のOMGのアーキテクチャーと実質同じものになる[3]。

最後に、マルチメディア間の相互変換技術は、文－音声変換（TSS；Text-To-Speech）や音声認識を含むが、これらの標準化（データフォーマット、APIなど）の動きは今のところはない。しかし、マルチメディアプラットホーム環境の整備が進むにしたがって、近い将来に標準化がおこなわれると考えられる。

6.3 音声処理技術標準化の将来

音声符号化の世界では、今後も一層の高能率を目指して標準化が検討されると考えられる。また、他のメディアとの同期を含めた標準化も必要になるであろう。

次に、現在進められているマルチメディア・プラットホームに関する一連の標準化が一段落すると、文ー音声変換や音声認識に関連した標準化が進められると考えられる。この場合、目的はアプリケーション構築者とデバイス（もしくはソフトウェアパッケージ）提供者の双方が、独立に開発をおこなえるオープン環境を目指したものとなるであろう。図6.3-1は、アプリケーション側とマルチメディア・プラットホームとを結ぶプロトコルとAPI規約を決めることにより、クロス・プラットホームでの互換性が保証されることを示している。このような枠組みは、国際的に統一されることが望ましい。

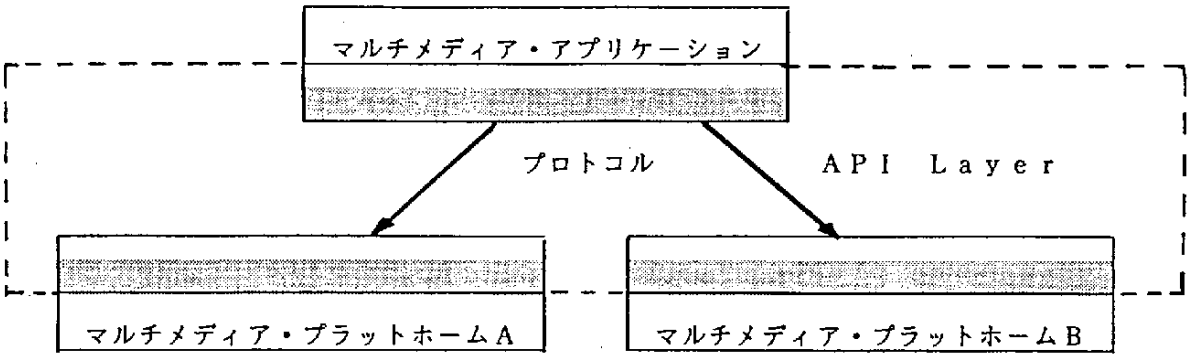


図6.3-1　API (Application Program Interface)

参考文献

[1] マルチメディアOS：日経エレクトロニクス, 1992年8月3日号, No.560, pp.93-112, 1992.8

[2] The Common Object Rquest Broker : Architecture and Specification, OMG, 1991.12

[3] Request for Technology : Multimedia System Services, IMA, 1992.10

第７章　今後の課題

第７章　今後の課題

第２部では、以上に述べてきたように、音声認識･理解技術を組み込むための支援ツール、評価手法、インタフェース条件の標準化、ユーザインタフェースの使用手順の標準化など応用開発支援関連の問題点に対応するために、いくつかの施策を提案した。

これらの提案をどのように実現して行くかが今後の大きな課題である。以下に各項目ごとに主な課題を述べる。

（１）音声認識･理解応用システムの分類型の普及

エキスパートシステムは、そのタイプを思い切って数種類の型に分類し、それを広く一般に広めることによって、応用イメージと技術課題が明確になり、その後の発展に結び付いたと言われている。今回の音声認識･理解応用システムの分類の型の提案も、この良き先例に習って、音声認識･理解応用システムの発展に貢献することが、その大きな狙いの１つとなっている。

したがって、この試案を広く学会等に発表し、多くの関係者の意見をいただき改良の上で、音声認識･理解アプリケーションシステムの開発関係者などにも広く周知することを計画していきたい。

（２）音声インタフェース開発支援ツールの具体化

画面情報を中心としたＧＵＩ（Graphical User Interface）は、ｍｏｔｉｆ TM[注1] やＯＰＥＮ　ＬＯＯＫ[注2] などが業界標準として作られ、数多くのＵｎｉｘ[注3] 応用システム開発の有力な開発ツールとなっている。

今回提案した音声インタフェース開発支援ツールなる概念は、音声インタフェースにおける標準ＧＵＩのような役割を期待しているものであり、ここで示した音声対話アプリケーション開発支援ツールは、その１つの具体的イメージである。このような視点から、音声認識･理解応用システムの開発支援ツールとして、その実現を図る必要がある。

しかし、より長期的に見るならば、音声だけでなく、映像や動画など他のメディアも取り扱え、標準ＧＵＩなどとの整合性も図られた、マルチ・モーダル・インタフェース・ビルダーのような、より広い概念のモノを、業界で策定し実現を図るべきであろう。

[注1] ｍｏｔｉｆ TMは、Ｏｐｅｎ　Ｓｏｆｔｗａｒｅ　Ｆｏｕｎｄａｔｉｏｎ　Ｉｎｃ．の登録商標です。

[注2] ＯＰＥＮ　ＬＯＯＫは、米国における米国ＵＮＩＸ　Ｓｙｓｔｅｍ　Ｌａｂｏｒａｔｏｒｉｅｓ，Ｉｎｃの登録商標です。

[注3] Ｕｎｉｘは，ＵＮＩＸ　Ｓｙｓｔｅｍ　Ｌａｂｏｒａｔｏｒｉｅｓ，Ｉｎｃが開発しライセンスしています。

（3）使用語彙選択支援ツールの具体化

当面音声認識・理解システムを製品化している各社で早急に用意することが、応用システム拡大の上で望ましい。長期的には上述の音声インタフェース開発支援ツールの一部として実現されることが必要になろう。

（4）評価手法の確立

アプリケーションシステム及び音声認識・理解システム双方の評価手法共に、実用現場からの実態を反映させ、改良を継続的に図る必要がある。システムや技術の発展や、応用の拡大に伴うユーザの意識の変化に伴い、その時々にあった評価手法に変化発展させていくことが重要であろう。

（5）標準化の提案、推進

本文にも述べたように、現在米国を中心に検討されているマルチメディア関連部のアプリケーションシステムへの組み込みインタフェースの標準化動向は、音声入出力もその対象として言葉だけは触れられているものの、その実態は信号（波形）としての音声の取り扱いが中心のように見える。

しかし、音声の利点を活かす立場から言えば、音声の強い実時間的性格と、さらに対話形式に見られる入出力間の有機的関連性を持った実時間的性格、言語情報としての記号性などを効果的に活かせるインタフェース仕様の検討と、その標準化を積極的に図る必要がある。仮に、これらの音声特有の性格を見落としたインタフェース仕様が、他分野からの提案に基づいて先行して作られ、広く普及する事態となれば、その内容によっては、音声の利点を活かす機会が将来大幅に制限される懸念が生まれる。

また、音声認識・理解システムのエンドユーザにとっての使い方（発声の方法、手順など）も各メーカ間で統一し、標準化を進めないと、音声認識・理解システムに対する期待感の高い高齢者や障害者にとって、非常に使い勝手の悪い、本末の顛倒したインタフェースとなる危険性が高い。しかし、音声インタフェースの使い勝手に関する検討はまだ明かでない点も多く、メーカ間で協力して開発を進める体制が作られることが望ましい。

なお、音声認識・理解技術自体の性能高度化、高度知的処理技術の開発等の技術に関する課題は、第3部で述べられる。

第３部　音声認識･理解技術の諸課題

第１章　概　要

第1章　概　要

1.1　はじめに

音声認識·理解技術の研究開発動向、技術課題については、昨年度報告書において詳しく論述してきた。しかし、その視点は技術開発側からのもであり、応用システム側からの技術ニーズの視点に立ったものではなかった。そこで、今年度は、将来の音声認識·理解技術の応用動向を予測して、応用システム側からの視点で必要となる要素技術、研究開発課題を考察することにする。

ここでは、おおむね10年後を目途に将来実現が見込まれる応用システムを類型化し、9種類の応用システムについて想定される利用形態や、必要となる技術課題について考察する。なお、音声認識·理解応用システムに必要な要素技術のうち、昨年度報告書に含まれてない技術の研究開発動向については、本報告書において昨年度同様に詳しく述べることにする。

そこで、第3部の構成は次のようになる。まず第2章では、昨年度報告書において扱わなかった4つの要素技術、

1）人工知能要素技術　　2）ユーザインタフェース
3）話者認識　　4）音声規則合成

について、その研究開発動向と技術課題を考察する。人工知能要素技術は、将来の高度な音声認識·理解システムにとってきわめて重要であるばかりでなく、音声認識·理解技術の応用範囲を拡大する上でも重要な役割を担うものと期待される。ユーザインタフェースは、利用者のユーザビリティを向上させる重要な技術であるが、これまで比較的研究が少なかった分野である。今後、拡大が見込まれるマルチメディア、マルチモーダル・インタフェースなどと併せて活発な研究開発が望まれる。話者認識技術は、多くの実際的な応用局面において利用者(話者)を限定したいという要求があり、必要となる技術である。音声規則合成は、音声応用システムにおいて重要な位置を占める音声応答機能のための基本技術である。

次の第3章では、応用システム別の技術課題について考察する。応用システムは以下の9種類である。

1）音声ワープロ　　2）福祉ロボット
3）ＣＡＩシステム　　4）情報案内システム
5）予約システム　　6）計算機用音声インタフェース
7）携帯端末　　8）マルチメディアデータベース検索システム
9）ナビゲーションシステム

これらの応用システムは、次節で述べるように、今後10年を見込んで考えられる音声認識・理解システムの応用を機能要素に分類し、バランスよく選択したものである。

1.2 応用システムの概要

音声認識・理解技術の応用を想定したとき考えられる基本的な機能として以下のようなものが挙げられる。

a）指示、命令（たとえば、ロボットに対する音声による指示など）
b）案内、問い合わせ（各種情報についての問い合わせ、データベースからの情報サービスなどに音声入力を利用）
c）予約（電話などの音声による各種の予約、注文など。案内との違いは、システム側のデータベースの書換えを伴う点）
d）文書作成（音声による定型文書や議事録などの作成）
e）教育、CAI（知的CAIへの音声入出力の組み込み、管制訓練、語学練習など）
f）自動通訳（音声理解、自動翻訳、音声合成を組み合わせたシステム）
g）汎用インタフェース（計算機に組み込まれた汎用の音声入出力）

本報告書では、これらの機能要素を含む典型的な応用システムとして、前節で挙げた9種類の応用システムを取り上げて考察する。ここでは、各システムの概要について簡単に述べておく。

1）音声ワープロ

音声を主入力機能とするワープロで、特定業務文書作成、自由文書作成、速記メモなどへの利用が考えられる。ユーザフレンドリなインタフェース、大語彙認識などの機能が望まれる。

2）福祉ロボット

高齢化社会の到来などを考慮すると、看護、介助等にロボットを利用する必要性が高まり、同時にこれらを音声により制御するという機能要請も増大すると考えられる。高齢者や身体の不自由な人などの多様な音声への適応が課題となる。

3）知的CAI（音声インタフェース機能）

知的CAIにとって音声言語インタフェースは必須である。利用者（学習者）は、システム側からの質問に対し思考による沈黙や質問、躊躇することなどが考えられるので、言い淀み、言い直しなどに対する言語処理が望まれる。

4）情報案内システム（マルチメディア）

計算機の高性能化によって、大量の情報を整理してデータベースとして蓄積することが可能になりつつある。この情報ベースから、音声を含むマルチモーダル入力・マルチメディア出力によって、ユーザが有益な情報提供を受けるシステムが考えられる。

5）予約システム

現在のところ、音声認識機能自体を予約に用いているシステムは見あたらないが、将来的には、発達した通信ネットワークを利用して、携帯端末や移動端末等からの音声による各種予約、買物注文などが考えられる。普及のためにはインフラストラクチャの整備、バーチャルリアリティ技術なども課題となる。

6）計算機用音声インタフェース

個人向けのパーソナルコンピュータやワークステーションの普及とそれらのマルチメディア機能化に伴って、音声インタフェースを組み込むことが考えられる。ここでは、アプリケーションプログラムとのインタフェースの標準化が重要な問題となる。

7）携帯端末

装置のハード面から考えると、音声入力を主体とするキーボードレスに近い仕様によって、携帯性のよい端末が考えられる。利用範囲は各種考えられるが、インフラ整備が重要となる。

8）マルチメディアデータベース（DB）検索システム

今後、普及増大すると見込まれるマルチメディアDBを音声で検索するシステムで、ユーザインタフェースに優れる利点がある。特に、音声DBファイルを直接音声で検索するシステムは、DB構築やDB管理システムの負担軽減につながる。

9）ナビゲーションシステム

自動車への音声認識関連技術の応用は、運転中、手が離せないという状況を考慮すると有望な分野の1つと考えられる。ここではその代表として音声インタフェース機能をもつナビゲーションシステムを取り上げる。

なお、自動通訳関連システムについては、すでに、ATR自動翻訳電話研究所のプロジェクトに関連して詳しく検討されてきているので、本報告書では取り上げないこととする。

1.3　応用システムからみた要素技術の研究課題

ここでは、昨年度報告書でその研究開発動向を詳述した音声認識・理解要素技術について、改めて簡単に復習し、これらの技術が将来の応用システムのどのような局面において重要となるかという視点から考察する。要素技術には、大語彙化のための認識技術、連続音声や自由発話、対話音声を扱う技術、話者適応化や不特定話者対応の技術、話者識別、発話環境（耐雑音、

電話音声、低品質音声）処理技術などがある。また、音声認識·理解システム関連技術として、音声合成技術やシステムのユーザインタフェース技術も重要である。

（1）大語彙音声認識

大語彙化は、一般的にはシステムの高度化や用途の拡大にとって欠かせない重要な要素である。また、システム構築において語彙の選択あるいは限定を無理なくおこなえること、利用者が語彙に関する制約を比較的意識せず利用できることなどの利点もある。しかし、一方では、認識率や応答時間などとは相反する関係にあり、応用上はシステムの目的によく適合した語彙量を選ぶ必要がある。

大語彙化のための技術課題としては、主に、音声特徴抽出技術、有効な音韻モデルなどの音響音声学的レベルでの高性能化が挙げられる。特徴抽出技術については、最近ではほとんどシステムが音声のフーリエ・パワースペクトル包絡を用いていて、新しい傾向が見られない。大語彙化に向けてはこの部分でのブレークスルーが望まれる。音韻モデルに関しては、隠れマルコフモデル（HMM）によるモデル化が多く、その精密化が認識率の改善に寄与することが明らかになっているものの、モデルのロバストネスとのトレードオフが問題となっている。また、大語彙化に伴う処理時間の増大を回避するための高速検索技術、有効な辞書構成と実装技術などが必要となる。

大語彙は多くのシステムにとって望まれる機能であるが、用途を限定したシステムでは、むしろ認識率や未登録語のリジェクト、応答時間が重視される場合もある。大語彙が基本的に必要な応用システムとしては、一般的な文書作成のための音声ワープロ、案内システムなどでは数字、名前など語彙数が多くなる対象を扱うシステム、あるいはDB検索で項目数が多いシステムなどである。逆に、福祉ロボットやナビゲーションのようなシステムでは、大語彙化よりも即応性のような応答時間、マクロ命令への対応などが望まれる。

（2）連続音声認識、自由発話音声、対話音声処理

発話の様式については、認識率の面からは孤立単語発声が断然有利であるが、ユーザインタフェースを考えると、連続音声さらには自由発話音声への対応が望まれる。このためには、従来の自然言語処理技術に加えて曖昧な入力を処理できる音声言語向きの言語処理技術が必要となる。また、自由発話音声処理には、音声に通常含まれる非言語音、笑い、言い直しなどの非文法的発話、未知語（辞書にあらかじめ登録されてない語彙）などの処理技術が要求される。さらに、音声言語にはしばしば、意味的誤り、多重性なども含まれる。これらの問題の処理は、自然言語処理の分野でも最も困難とされてきたもので、音声との結合ではさらに複雑となる。ただし、これらの言語現象については、自然言語処理の面のみでなく、韻律など音声的側面か

らの研究も今後進めていくべき研究課題である。

自由発話音声や対話音声については、人工知能技術、特に対話モデルなどとの関連も強く、後の第2.1節における技術動向解説を参照されたい。これら課題についての音声研究の歴史は浅く、蓄積されている知識も少ない。有効な音声データベースの構築が当面の課題の1つである。最近の傾向として、統計量にもとづいた確率言語モデルなどの提案もあるが、そのために必要となるデータ量は莫大であり、現実的な解決をみたものは見あたらない。このためには、音声言語データの計量化法、自己組織化や構造化法、タスクの複雑さの評価法、音声理解システムの性能評価法等についての研究も重要である。

上述したように、自由発話を許容することはユーザインタフェースの大幅な向上をもたらす。特に、案内やコンサルティング、予約システム、知的CAI等への応用についてはその実現が強く望まれる。しかし、認識性能、システム構築上の難度を考慮すると、自由発話の許容は不利であり、システムの確実、安定な動作という点では不安がある。一方、孤立単語発声や、文節単位発声でも使用上の著しい不都合はないシステムも考えられる。たとえば、音声ワープロ、単純な案内、指示・命令を主とするロボットのようなシステムである。

対話型のシステムは単純なものから高度なものまで考えられるので、対話音声の範囲はかなり広くなる。ロボットのような指示命令機能の場合には、単純な対話でも有用な場合が多い。これに対して、知的CAIのようなシステムでは、利用者の反応によってどのような応答を返すか等の指導方略に関連してきわめて高度な音声対話処理が考えられる。

(3) 話者適応、不特定話者音声認識

音声認識装置が任意の話者の発話を何の訓練もなしに認識できれば、個々の話者の声を登録する方式に比べ、明らかに利便性に優れる。しかし、音声の音響物理的特性は同じ音声言語でも話者によって多様に変化するため、いわゆる不特定話者音声認識はきわめて困難な技術の1つである。次善の方式として、少ない音声サンプルを用いて、発話者の特性にシステムの適応化を図る話者適応化技術があり、最近活発に研究されている。話者適応化処理は、システムの性能を特定話者方式の性能とほぼ同等に引き上げることができ、応用面から現実的な方策の1つである。しかし、話者がきわめて多数でかつ適応化処理のような余裕が認められない類の応用も数多く考えられる。このような場合には、必然的に不特定話者音声を対象とするシステムが必要になる。

話者適応化技術の多くは、標準話者と入力話者とのベクトル符号の対応付け、特徴空間分布のパラメータの対応付けにもとづいている。いづれも静的ないし局所的な音韻的特徴の適応であり、より動的あるいは発話形態的特徴の適応などは今後の課題である。

不特定話者化の技術は、音韻表現モデルに若干の違いはあっても基本的には複数標準パター

ン化である。これは方式上、現実的な選択であるといえるが、任意の年齢や地域など広範な話者を考慮した場合、システムのロバストネスの点で依然として大きな問題がある。

話者適応化技術は、いずれ将来的には、いわゆる特定話者方式を包含するものになると考えられる。したがって、主に個人向け仕様のシステムや、利用者が少数に限定されるようなシステムに組み込まれるであろう。たとえば、音声ワープロや端末、ロボットなど。しかし、多くの不特定利用者の随時使用を想定するシステムでは、不特定話者音声認識が必要とされる。いわゆる学習によって短時間のうちに話者に意識させることなく適応化するというシステムは方式としては考えられるが、応用を想定すると実用性が低いように思われる。

（4）話者認識

音声分析にもとづく話者認識には、個人（話者）の同定（照合）と話者識別があり、おおむね、指紋などと同じようにセキュリティへの応用の観点から研究されてきた。この機能は、音声を利用するというユーザフレンドリ性に優れ、多くの分野で応用可能性をもっているので研究はかなり長期にわたりおこなわれてきた。音声認識・理解装置への組み込みに関しては、装置が特定の利用者にのみ使用可能であるようにするという要請からきている。特に、装置側がリアルタイムに特定の話者のみの声に選別的に反応するならば、その効果はきわめて大きいといえる。

話者認識技術の現状と課題については、第2.3節で解説されるので、参照されたい。話者認識技術には、あらかじめ話者の話す言葉を知っている場合（テキスト依存）と、知らない場合（テキスト独立がある）。上記で述べたような使用状況を想定した場合は後者が要求されることが多く、しかもリアルタイム性を要求されると、現状の技術ではきわめて困難といえる。また、同じ個人であっても一般に時期（時間）や、健康状態、周囲環境による人間の声の不安定さは、よく知られるところである。機能としての利点は明かであっても、実現についての技術的困難も大きい。

（5）発話環境処理

各種の雑音、騒音環境など音声入力における発話環境処理は、実使用を想定するときわめて重要な要素技術である。騒音の種類は使用環境に依存して変わるが、使用環境を決定する要因として次のようなものが考えられる。まず装置の設置環境である。これには電話系や、携帯端末を利用するというものも含まれる。従来は、工場、作業場などでの使用が多かったが、今後の応用を見込むと家庭内、公共スペース、オフィス、自動車内などの環境での使用が増大すると考えられる。

雑音処理技術は、大きく分けて、マイクロホンなどの集音装置に工夫をする方式とソフト的

に信号処理技術により雑音低減をおこなう方式がある。前者にはたとえばノイズキャンセリングマイクやマルチマイクの使用がある。後者としても様々な手法が提案されているが、雑音の性質が変化し易いものや非定常な雑音に対しては有効な手法の開発はきわめて困難である。

使用環境には、ハード的な側面とソフト的な側面からの問題がある。ハード面とはマイクと発話者との相対的な位置関係などで、マイクの種類もこれによってある程度限定されてくる。ヘッドセットのマイクは耐雑音性では有利であるが、一般には敬遠されることも多く、公共的な使用には適さない。また、ロボットのような場合、移動するロボット自身にマイクがセットされるとすれば、処理系はきわめて困難な問題となる。ソフト面では、いわゆるロンバート効果のように、使用者（話者）が使用環境に影響されて、発話の変動が大きくなるという問題がある。この問題は、環境が特殊になればより大きく出てくる。たとえば、ささやき音声なども使用環境によっては想定する必要がある。

外部騒音や回線雑音のような雑音音声とは異なるが、高齢者や傷病者、障害を持つ人の音声処理も将来の応用を想定すると重要な課題である。

また、装置自体の耐性などの面での耐環境性、装置の仕様とその他設備、法規、社会環境との整合性などももちろん実際上は問題となる。

（6）音声合成

音声認識･理解機能を組み込んだシステムは、多くの場合、音声応答機能を有している。システム側からの応答は、一般には入力自体についての確認と入力に対する回答や動作に分かれるが、そのメディアとしては用途に応じて様々なものが考えられる。音声応答は音声が応答として便利ないし効率的な場合についての使用はもちろんのこと、システムの親しみ易さの増大にも貢献する。これらの音声応答は音声規則合成技術によって実現されている。音声規則合成は、入力として日本語文章テキストを想定した「テキスト-音声変換」の機能とも言えるが、応答の内容を表す概念を出発点と考えてもよく、この場合は「概念からの音声合成」と呼ばれる。

音声規則合成技術における問題点は、1）テキスト解析などの言語処理技術が不十分であること、2）韻律（アクセントやイントネーションなど）の自然性の確保、3）音質の自然性、肉声感の確保、などにある。最近の規則合成方式には、大量の音声素片を蓄積しておきその接続にもとづく方式が多く、音質の自然性などの面でかなり進歩してきている。しかし、音質自身の制御、韻律の制御など本来パラメータによって制御すべき要素については問題解決に至った訳ではない。音声の技術動向については第2.4節で詳述する。

音声合成は、応用面からみた総合的評価法の確立が重要であり、最近ようやくその観点からの研究が開始されつつある。すなわち、基本的な評価要素である明瞭性や自然性についても、

場面や状況に応じてどのような音声が有効であるのかといった観点からの評価が重要である。したがって、単一（同質の）の音声を合成する手法ではなく、パラメータによって声質の制御ができる方式が将来望まれる方向である。

（7）ユーザインタフェース

システムの総合的評価はユーザインタフェースの良否に大きく依存するので、ユーザインタフェースの重要性は論を待たないが、比較的定量的な調査研究結果を出しにくい分野である。このため、既存の技術や知見を踏まえて次の技術が登場するという側面が相対的に少ない。ユーザインタフェースを構成する技術要素には、表層的な操作性に依存する部分と、より深いたとえば人工知能技術や音声処理技術の本質的な進歩に依存する部分があり、それらがある時点で統合化されて1つのシステムとなる。

ユーザインタフェースの1つの切口としてモーダリティがある。将来的には明らかにマルチメディアを用いた情報提供とマルチモーダルな入力が優勢になると予想される。しかし、このようなマルチモーダルという点の有効性・有用性を明解に示した実験は必ずしも多くなく、今後の研究に負うところが大きい。これらの技術動向については第2.2節で詳述される。

ユーザインタフェースの向上のためには、ユーザのモデル化が重要な課題となる。将来的には、ユーザの応答から音響音声的情報を利用してユーザの感情などを推測するような情緒情報処理なども考えられる。また、ユーザインタフェースの評価法も重要な課題の1つである。

第2章　要　素　技　術

第2章　要素技術

2.1　人工知能要素技術

2.1.1　はじめに

本節では、人工知能研究の成果をいくつか紹介し、これらの成果を音声処理にいかに利用できるかについて述べる。

音響的な情報だけから音声を正しく認識することは一般に困難である。実際、我々人間も完全に音が聞きとれなくても、相手の発話を理解することができる。これは、我々が音声を認識するときに、音響的な情報だけでなく言語的な知識や常識を使って音響的に不足している情報を補っているから可能なのである。

本節の前半では、音声処理に利用できる、より高次の知識学的な知識について述べる。まず、2.1.2項では、辞書、文法などの統語的知識について、その記述法、それを用いた統語解析の手法について紹介する。2.1.3項では、意味的な知識について、2.1.4項では、実際に言語を使って情報のやりとりをする際の様々な制約に関する語用論的な知識について述べる。

本節の後半では、人工知能の分野で提案されたいくつかの処理手法について述べ、その音声処理への応用例をいくつか紹介する。まず、2.1.5項では、黒板モデルについて述べる。黒板モデルはARPA(DARPA)のプロジェクトであるHEARSAY-II［Reddy 77］で用いられたことで有名である。2.1.6項では、同じくARPAプロジェクトで開発されたHARPYシステム［Lowerre 76］で採用されたネットワークモデルについて述べる。2.1.7項では、最近の音声言語理解システムで広く用いられている、ネットワークのノード間でマーカと呼ばれるデータをやりとりすることによって処理をおこなうマーカパッシングと呼ばれる手法について説明する。

2.1.2　統語的知識

統語的知識は辞書と文法に大きくわけることができる。辞書は単語がどのようなアルファベットの系列からできているかという知識を与える。一方、文法は単語と単語の連接関係、あるいは単語列が構成する句や節などのより大きな構造を規定する知識である。

（1）　辞　書

音韻などを認識の単位とするシステムでは、辞書は音素がどのように結合できるかを規定する。効率のよい辞書のデータ構造として、TRIE構造［Bahl 89］やネットワーク構造［Lleida 92］などが知られている。TRIE構造の辞書では、図2.1-1に示すように、各弧にアルファベットを付与した木構造で表現する。この例では、/k o k o/とk o r e の接頭部/k o/が共有さ

れている。同じ接頭部を持つ単語はノードを共有することができるので記憶空間を節約できるとともに効率のよい検索が可能になる。

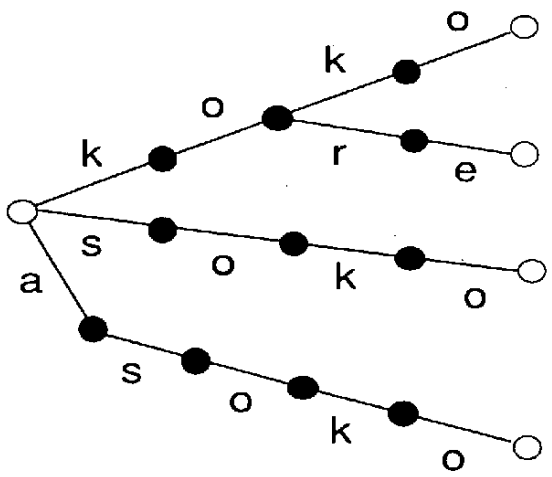


図2.1-1　TRIE 構造の辞書の例

ノードの共有を接尾部についてもおこなうとネットワーク構造の辞書ができる。図2.1-2に例を示す。この例では、/s o k o/と/a s o k o/の接尾部/o k o/が共有され、さらにこの2つと/k o k o/の最後の/o/の部分が共有されている。

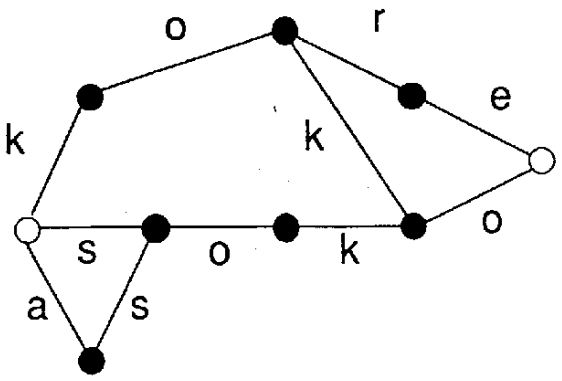


図2.1-2　ネットワーク構造の辞書の例

（2）文　法

文法は単語と単語の連接関係、あるいは単語列が構成するより大きな構造を規定する。もっとも単純な文法として文テンプレート（Sentence Templete）や単語対文法（Word Pair Gram-mar）［Lee 89］などがある。文テンプレートは、文として許される単語の範疇や単語の系列を１文単位に作成したものである。図２.１-３に文テンプレートの例を示す。この図では２つの文が定義されている。図中で＜＞で囲まれているのは単語の範疇を、それ以外は単語を表す。

⟨what-is⟩ ⟨optthe⟩ ⟨shipname's⟩ ⟨gross-ave⟩ displacement in ⟨long-metric⟩ tons
Is ⟨optthe⟩ ⟨shipname's⟩ ⟨earliest⟩ ⟨casrep⟩ rated worse than ⟨optthe⟩ hers

図２.１-３　文テンプレートの例

単語対文法は、単語と単語の隣接関係をによって文法を記述する。これらの文法はいずれもチョムスキーの階層でいう正規文法のクラスに属し、有限状態オートマトンで表現することができる。このため、これらの文法の統語解析器はそのままDPの制御に用いることができるという利点がある。また、認識の対象となる文の集合が比較的小さく、閉じているような場合には容易に文法を作成することができる。しかしながら、自然言語を正規文法で扱うには、表現力の点で限界があるので、複雑なタスクを実現するのに正規文法を用いるのは現実的ではない。

文脈自由文法は正規文法を真に含む文法クラスであり、効率的な統語解析手法が知られていることから、自然言語処理の分野ではもっとも一般的な文法クラスである。文脈自由文法の統語解析手法としては、ATNG［Woods 73］、チャート法［Kay 80］、LR 法［Aho 72］［Tomita 87］などがよく用いられる。

ATNG（Augmented Transition Network）：　ATNG は文脈自由文法をネットワーク形式で表現する文法形式で、開始ノードからネットワークをたどることによって、トップダウン縦型探索で統語解析をおこなうことができる。たとえば、図２.１-４の文法は図２.１-５のようなネットワークで表現できる。

(1)	S → NP VP	(5)	ADJP → adj
(2)	NP → ADJP noun	(6)	ADJP → adj ADJP
(3)	NP → det ADJP noun	(7)	VP → vi
(4)	NP → pronoun	(8)	VP → vt NP

図２．１-４　文脈自由文法の例

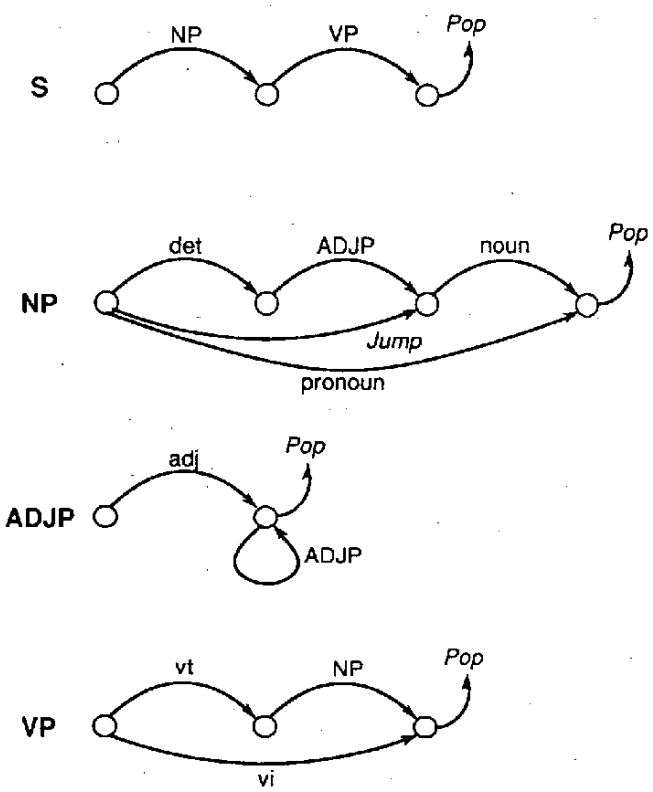


図２．１-５　図2.1-4の文法のATNG表現

ネットワーク中の状態遷移はその状態から出ていく弧に付与された文法範疇を手がかりにおこなう。文法範疇が非終端記号（図中では大文字のシンボル）であれば、その非終端記号のネ

ットワークに遷移する。このようにATNGはプッシュダウンオートマトンと等価であることがわかる。ATNG では、さらにネットワークの各弧にその弧を通過するときにおこなうアクションを記述できる。このアクションによってATNG が持つ種々のレジスタに値を設定したり、レジスタの値を検査したりすることができる。この機構を用いて人称や数の一致の検査を容易に記述することができる。この点でATNGの能力は純粋な文脈自由文法を越えている。

チャート法： ATNG の解析アルゴリズムは基本的に縦型探索なので、バックトラックにより再計算を繰り返し、最悪の場合、文の長さに対し指数オーダの解析時間が必要となる。チャート法は解析途中でできた部分構造を保存しておき、同じ部分列に対して再計算を繰り返さないように工夫した横型の統語解析手法である。チャート法はトップダウン、ボトムアップのいずれの方法も利用できる。図２．１-４の文法を用い、"I have a pen"をボトムアップに解析したときにできるチャートを図２．１-６に示す。

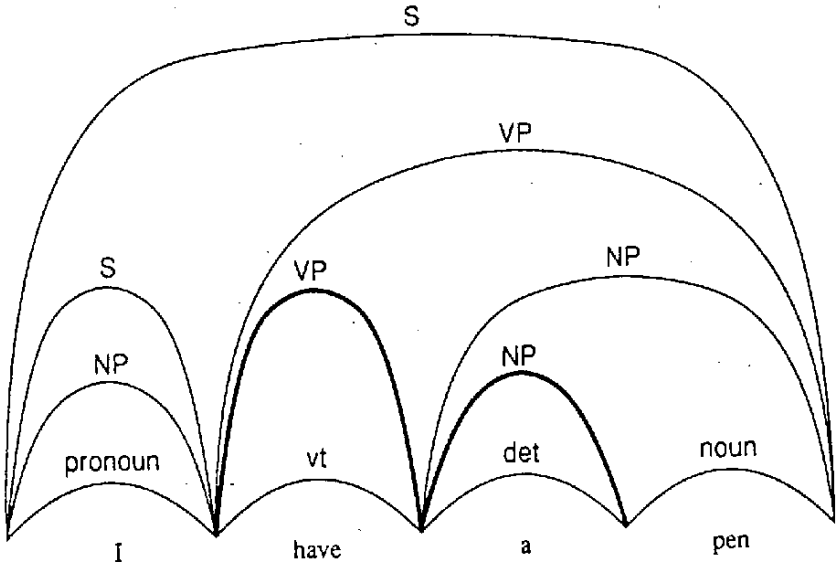


図２．１-６　チャート法による解析例

図２．１-６において、非終端記号あるいは語彙記号（小文字で表記）をラベルに持つ弧の集合をチャートと呼ぶ。弧には完成した部分構造に対応する不活性弧（細線で表示）と不完全な部分構造に対応する活性弧（太線で表示）がある。不活性弧でおおわれた文字列がその弧のラベルで示される部分構造にまとめられたことを表している。解析はこの図に弧を追加することによっておこなう。最終的に入力文字列全体を開始記号（この場合はS）でラベル付けされた

不活性弧がおおえば解析は成功したことになる。チャート法を音声処理に利用した例として岡田らの研究［岡田 90］がある。

LR法： LR法は文脈自由文法の部分クラスであるLR(k)文法を決定的に処理できる手法であり、与えられた文法からあらかじめ作成したLR表を参照して、現在の状態と次の入力シンボルからアクションを決定する。可能なアクションとしては、シフトとレデュースがある。シフトは入力シンボルをスタック上に積む動作であり、レデュースはスタック上のいくつかのシンボルをポップして、それらのシンボルから構成できるシンボルをスタック上にプッシュする動作である。レデュースによってチャート法でいえば不活性弧ができることになる。

図２.１-７に図２.１-４の文法から得られるLR表を示す。

	action 部							goto 部			
	det	noun	pronoun	adj	vi	vt	$	S	NP	ADJP	VP
0	sh4		sh5	sh6				1	2	3	
1							acc				
2					sh8	sh9					7
3		sh10									
4				sh6						11	
5					re4	re4	re4				
6		re5		sh6						12	
7							re1				
8							re7				
9	sh4		sh5	sh6					13	3	
10					re2	re2	re2				
11		sh14									
12		re6									
13							re8				
14					re3	re3	re3				

図２.１-７　図2.1-4の文法から得られるLR表

LR表はaction部とgoto部からなっている。LR表の各行は各状態に対応し、各列は、action部では次の入力シンボル、goto部では各非終端記号に対応する。action部は、現在の状態(行)と次の入力シンボル(列)から次のアクション（シフトまたはレデュース）を規定する。ここで、reNは規則Nを用いてレデュースをおこなうことを、shNはシフトして状態Nへ遷移することを示している。goto部は、レデュースによって新しいシンボルが完成したときに次にどの状態へ遷移するかを規定している。

LR(k)文法は文脈自由文法の部分クラスなので、一般に文脈自由文法からLR表を作成するとある状態と次入力で複数のアクションが可能になる場合がある。これをコンフリクトという。LR表でいえば、表中のある項目に複数のアクションが書きこまれることになる。富田はコンフリクトのあるLR表についても効率的に統語解析がおこなえる一般化LR法と呼ばれる手法を提案している［Tomita 87］。音声処理においても一般化LR 法が利用されている［北 90][斎藤 88］。

確率文脈自由文法： 文法のクラスを正規文法から文脈自由文法にすると一般に複数の解析結果が得られることが多くなる。この中から正しい解析を選択することをあいまい性の解消と呼ぶ。あいまい性の解消は自然言語処理の中心的な研究テーマの一つであり、これまでにも多くの研究がなされている［長尾 92］。しかし、音声入力の場合には、音声処理の結果にも多くの候補が存在するので入力自身にもあいまい性を含むことになり、書き言葉の解析にくらべ状況はさらに厳しくなる。音声処理においてあいまい性を扱う手段として、確率文脈自由文法を使うことが多い。確率文脈自由文法は、通常の文脈自由文法の各生成規則に、その規則が使用される確率を付与したものである［Wetherell 80］。たとえば、図２.１-８に確率文脈自由文法の例を示す。

(1)	S → NP VP,	1.0	(5)	REL → pronoun VP,	0.3
(2)	NP → proper-noun,	0.4	(6)	VP → vi,	0.5
(3)	NP → det noun REL,	0.6	(7)	VP → vt NP,	0.5
(4)	REL → ε,	0.7			

図２.１-８　確率文脈自由文法の例

ここで、大文字のシンボルは非終端記号を、小文字のシンボルは終端記号を表す。確率は同じ非終端記号を左辺に持つ規則の和が１になるようにふられる。たとえば、規則(2)と(3)はいずれも名詞句(NP)を展開する規則であるが、これは名詞句が固有名詞（proper-noun）に展開される確率が0.4、冠詞(det)、名詞（noun）、関係節（REL）に展開される確率が0.6であると解釈できる。解析結果の尤度は、その解析結果を導出するのに使用された規則に割りふられ

ている確率の積で計算することが多い。ただし、この方法だと、多くの規則を使用した解析結果の方が尤度が低くなるので、規則数で幾何平均をとる場合もある。

個々の規則の確率は訓練データを用いて学習する。確率の学習法として、ベイズ推定に基づく藤崎らの手法［Fujisaki 91］やInside-Outsideアルゴリズム［Lari 90］などが知られている。Inside-OutsideアルゴリズムはHMM（Hidden Markov Model）［Rabiner 86］の学習法であるBaum-Welchアルゴリズムを文脈自由文法用に拡張したものである[1]。

図2.1-8の例をみてもわかるように各規則にふられた確率は同じ非終端記号を左辺に持つ規則を根元事象とする局所的で静的なものである。したがって、前述のように、導出に使用された規則の確率の積が必ずしも大域的な尤度を与えるとは限らない。どのような文脈で規則が使われるかによって、各規則の確率は変化すると考える方が自然であろう。このような考えから、確率文脈自由文法の確率に文脈依存性を導入する研究もおこなわれている［Su 91］［Kita 92］［リグレー 92］。これらはいずれもある非終端記号が展開される時に、その前後にどのような(非)終端記号が存在するかを考慮している。これらはN-gramモデルを確率文脈自由文法に導入したものと考えることができる。

確率文脈自由文法の学習には例文と、例文から導出されるすべての解析結果を用いることが多い。このような訓練データを用いて規則の確率を適切に学習すれば確かにperplexityは下がるが、必ずしも正しい解析の尤度が高くなるという保証はない。Pereiraらは訓練データに正しい構造だけを使うことによってより精密な確率の学習がおこなえることを示している［Pereira 92］。また、Inside-Outsideアルゴリズムの計算量は一般に$O(n^3)$であるが、完全に正しい解析だけを使えば$O(n)$で学習ができるという利点もある。しかし、一般に正しい解析だけを含むような訓練データを用意するのはコストがかかるし、文法を修正するとこれらのデータが使えなくなったりするので、現実的にはこれは困難であろう。

正規文法が有限状態オートマトンで実現できるのに対し、文脈自由文法はプッシュダウンオートマトンで実現できる。プッシュダウンオートマトンで直接DPマッチングを制御するのは困難なので、文脈自由文法を使う場合には何らかの工夫が必要になる。Pereiraらは、文脈自由文法を正規文法で近似する手法を提案している［Pereira 91］。この手法を使えば文脈自由文法で記述された文法を有限状態オートマトンで近似し、これを直接DPマッチングの制御に使うことができる。しかしながら、この方法では、文法が複雑になったり、語彙が大規模になると有限状態オートマトンの状態数が爆発するという問題もある。

---

[1] HMMは正規文法に対応することに注意。

2.1.3 意味的知識

音声認識で生じる膨大な数の仮説を意味的知識を用いて制限する手法がいくつか考えられている。その場合で、よく用いられる意味的知識の形式として格フレーム（case frame）がある。

格フレームは、主として動詞に関する知識表現の枠組であり、動詞の主格や目的格としてどのような意味の名詞がくるかを記述したものである。たとえば、「読む」という動詞は、主格に「人間」が、目的格に「字が書かれたもの」がくることが予測でき、「飲む」という動詞は、主格に「動物」が、目的格に「液体」がくることが予測できる。さらに主格や目的格が動詞に対してどのような意味的役割を果たすかについても記述される。「読む」と「飲む」の主格は行為者（agent）を表し、目的格は行為の対象（object）を表す。このような主格や目的格の行為に対して果たす意味的役割は、格文法では深層（deep case）と呼ばれる。

以下に格フレームの例を示す.

```
(def-frame 飲む
  (is-a 行為)
  ;; 上位概念に「行為」がある.
  (:agent (主格動物))
  ;; 行為者(agent) は動物である.
  (:object (目的格液体))
  ;; 行為の対象(object) は液体である.
)
```

構文解析によって文中の動詞や名詞句の位置などを知ることができ、名詞句についた格助詞などの情報から、その名詞句が主格や目的格になり得るかを知ることができる。そのため、格フレームに文を当てはめることにより意味解析をおこなうことができる。たとえば、「太郎がビールを飲む。」という文から、構文解析によって[S [NP 太郎が] [VP [NP ビールを] [V 飲む]]] という構造が得られたとする。この構造から動詞「飲む」を抽出し、それに対応する格フレームを用意する。次に「ビールを飲む」の構造と格助詞「を」から、「ビール」が目的格であることが分かり、先の格フレームのobjectスロットに入るかどうか決定する。これは、「ビール」に対して(ビール(is-a 液体))のような知識があれば決定できる。このとき、スロットへの埋め込みは成功する。同様に、「太郎」が主格であることが分かり、この格フレームのagentスロットへの埋め込みは(太郎(is-a 人間)) (人間(is-a 動物)) のような知識があれば成功する。このような意味的な整合性の検査をおこなうことを、選択制限(selectional restriction)

の適用と呼ぶ。

これから分かることは、まず、意味解析のためには概念間の上位/下位関係を表すシソーラスが必要であるということと、格フレームを用いることにより選択制限による解釈の候補の絞り込みをすることができるということである。ただし、一般に文の意味的妥当性を決定するためには膨大な背景知識が必要とされる。たとえば、選択制限では比喩的表現の意味的妥当性を判断することは非常に困難である。そのため、格フレームを用いた意味解析で、選択制限を違反した場合の解釈を排除せず、意味解釈に対するスコアとして反映させるべきだとする考えもある。このスコアの意味論として先ほどの確率的構文解析のように主観確率論を基盤とする考え方も現れているが、意味的知識をすべて確率論的に定義することは不可能なことであろう。

2.1.4 語用論的知識

意味的知識を用いても解釈が特定できない場合は多い。そこで、常識的知識や対象領域・文脈の情報を有効に利用する必要がある。たとえば、日常生活において繰り返し経験する定型的な行動パターンを構造化して蓄える記述形式としてスクリプト（script）［Schank 77］と呼ばれるものがある。

人は日常生活の様々な場面に関して多くの知識を持っている。つまり、レストランに行ったり、医者にかかったりする場面がどのような事象から構成されるか、またそのような場面において人が典型的にどのような行動をとるか知っているということである。このような知識を形式化して曖昧な解釈を特定化するという目的のために、スクリプトが考案された。スクリプトは演劇の台本のように、場面を単位として知識をモジュール化する。たとえば、「レストランで食事をする」スクリプトは次のように構成される。

スクリプト: レストラン

登場人物: 客A, ウェイトレスB, レジC

道具: テーブル、 メニュー、 料理, ...

初期状態: A は空腹, A はお金を所持, ...

最終状態: A は満腹, A のお金は減少, ...

事象:

1. A がレストランに行く.
2. B がA をテーブルに案内する.
3. A がテーブルにつく.
4. B がA にメニューを渡す.
5. A がメニューを見る.

6．A がB に料理を注文する．

7．B がA に料理を運ぶ．

8．A が料理を食べる．

9．A がC にお金を払う．

10．A がレストランを出る．

場面に関する知識をスクリプトの形で蓄えることにより、次のような文章を意味的に関連づけることができる。「レストランに行って、ステーキを食べた。このとき、財布がないのに気づいた。」第一の文より、レストランのスクリプトが想起され、また料理の種類が特定される。次の文で、支払いの場面において、日常とは異なった事態が発生したことが分かる。レストランに行く前には「空腹であった」ことや「メニューを見て、ウェイトレスに注文した」ことなどは、レストランスクリプトを参照することによって必要に応じて推論される。また、逆に日常的でない事態の発生に関しては、文章中に陽に述べられていることが必要になる。

しかし、日常生活を記述した文章の理解において、記述される場面のスクリプトをあらかじめ用意しておけるほどには定式化できない場合は多い。そこで典型的な場面を記述するスクリプトとは別に、人の目的（ゴール）とそのときにとる行動（プラン）に関する知識を蓄える必要がある。利用できるスクリプトがない場合でも、このようなゴールとプランに関する知識を用いて文章に含まれる文の意味的な関連づけをおこなうことができる。

また、たとえば病院に行くスクリプトと美容院に行くスクリプトでは、待合室の場面と支払いの場面は多くの共通点を持つ。そこで、スクリプトにおける共通部分を「オフィス訪問」、「契約」、「支払い」などのように抽象化することが考えられる。場面に関する大量の知識も抽象化し、階層化して蓄え、必要に応じてスクリプトに具体化すれば知識のためのメモリを大幅に節約することができる。Schankは、このような知識の構造を動的に構成される記憶モデルとして形式化し、MOP（Memory Organization Packet）［Schank 82］と呼んだ。MOPは単に知識を効率良く蓄えるための構造としてではなく、文章理解の状態の表現にも用いられる。

このような語用論的知識は、対話システムにおいて対象となるタスクを記述したタスクモデルとして用いることができる。また、ユーザの適性に合った対話システムを実現するためには、システムがユーザに関する知識を持たなければならない。たとえば、ユーザがあるタスクに関して初心者なのか、熟練者なのかによって適切な応答は違ってくるべきであろう。このようなユーザに関する知識はユーザモデルと呼ばれる。ここでは、ユーザモデルの表現法について、これまでに提案された手法を簡単に紹介する。

1) ステレオタイプモデル

実際的な方法として、ステレオタイプ型（stereotype）の知識を利用する方法が従来から考えられている。たとえば、推薦図書を選定するシステムGrudy［Rich 79］がある。以下に、このシステムにおけるステレオタイプモデルの例を示す。

SPORTS-PERSON

Triggers

(used-description 'athletic')

Facets

Name Value Rating

Motivations excitement 600

Interests sports 800

Thrill 5 700

Generalizations

ANY-PERSON {a canonical user model}

ステレオタイプ型ユーザモデルは、'body'と'triggers'と'hierarchical relation'の3 要素からなり、bodyはステレオタイプを適用するユーザの典型的情報を、triggersはbodyに記された値や固有の入力などの集合として記述し、hierarchical relationはほかのステレオタイプへ至る階層的関係を表す。

if-then 型の規則を使ってステレオタイプを記述することもできる。

if goal=plan(business-trip), then subgoal=get-flight(business-class)

if goal=plan(pleasure-trip), then subgoal=get-flight(economy-class)

2) プラン・ゴールモデル

ステレオタイプ型知識を用いたユーザモデルでは、システムは話し手のプランやゴールに関する信念を理解しない。そのため、ユーザのプランとゴールを考慮するユーザモデルの構築法が研究されている［Carberry 90］。このアプローチでは、新しい発話が与えられるごとに、次のようにしてユーザモデルを構築していく。

・発話をシステムのドメイン知識と関連づけ、ドメイン依存のサブゴールと関連したサブプラン（それは話者の注意がおかれているタスクを表現している）に関する仮説を立てる。

・これらの潜在的サブゴールとサブプランを、ユーザの全体のプランに関するシステムの信念と関連づけ、彼の注意の中心となる特定のサブゴールとサブプランを決定する。
・すでに推論されたプランを、新しいサブゴールやサブプランを含むように適切に拡張する。

なお、タスクモデルとユーザモデルの対話システムへの具体的な適用例については、本報告書の第2部第3章を参照されたい。

2.1.5 黒板モデル

これまでに述べてきた種々の知識をどのように組み合わせて用いるかに関して、いくつかの制御モデルが提案されている。ここでは、探索空間の設定と知識源の起動の観点からモデル化した、黒板モデル（blackboard model）と呼ばれるものを紹介する。

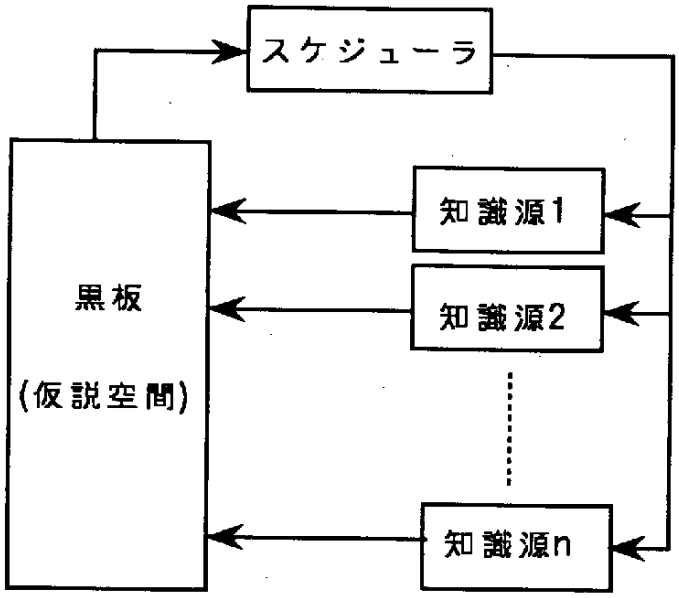


図2.1-9 黒板モデル

図2.1-9に黒板モデルの構成を示す。探索空間としては仮説空間全体をとり、これを黒板と呼ぶ。黒板上の仮説は、処理レベルに関係なく一様な構造を持ち、統一的な方法で評価値が与えられる。また、どの知識モジュールからも共通にアクセスできるようになっている。各知識モジュールは自分が動作可能になるための前提条件を持ち、黒板上にこの前提条件を満たすものが現れたときに動作可能となり、黒板上の仮説を検証してその評価値を設定したり修正したり、また自己の生成した新しい仮説を黒板に書き込むことができる。このように、各知識モ

ジュールは黒板との間で仮説を通して通信するので、モジュール間に従属関係がなく、独立に動作できる。各知識モジュールは並列に動作しながら、互いに仮説の生成と評価を繰り返し、最終的に発話全体を説明する文仮説を生成するまで処理を実行する。

黒板モデルでは、黒板上の仮説の構造と評価を処理レベルと独立に一様にしなければならないことや、知識モジュールを並列に動作させると無駄な処理が増大し、効率が著しく悪くなることから、動作可能な知識モジュールをどのようにスケジュールするかなどの困難な問題が存在する。

黒板モデルを最初に実装した音声理解システムHEARSAY-II［Reddy 77］では、計算機のパワーの問題から並列処理ができなかったため、各知識モジュールが独立に生成する仮説をスケジューラの制御に基づいて整理することで音声処理をおこなっていた。構文解析としては、最も確からしい単語や文節を中心に左右両方向に解析を進めていく島駆動法（island-driven method）が用いられている。

2.1.6　ネットワークモデル

図2.1-10にネットワークモデルの例を示す。このモデルでは、単語・構文・意味などの知識を、あらかじめ特徴パラメータの時系列または音韻系列な状態遷移ネットワークで記述し、音響処理部が生成するラティス状の仮説のあるパスがネットワークのどのパスと良く整合するかを探索する。ビーム探索法（beam search method）を用いて照合のスコアの低いものを枝刈りしながら、仮説とネットワークの間のマッチングをおこない、最適経路に相当する最適な文仮説を生成する。

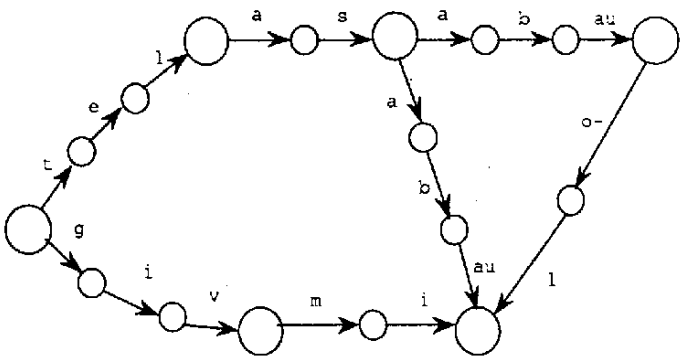


図2.1-10　ネットワークモデル

ネットワークモデルに基づいて設計開発されたHARPYシステム［Lowerre 76］が、単語数

1000程度で比較的単純な構造の文を対象として、成功率95%という良好な成績を残している。音声認識の重要な技術として発展したベクトル量子化とHMM（Hidden Markov Model）はネットワークモデルと相性が良く、1980年代に研究されている音声理解システムでは、このネットワークモデルがよく利用されている。この場合のネットワークは、音韻グラフ中の各音韻にそれぞれに対応するHMMを埋め込んだ形となる。

ネットワークモデルでは、トップダウン的な知識源はすべてあらかじめコンパイルされてネットワークに組み込まれている。音声理解の処理過程では、このトップダウン的仮説とボトムアップ的仮説(たとえば、音響分析結果のベクトル量子コード) の照合作業が大部分を占めるため、計算効率は非常に高い。しかし、知識が静的に統合されているため文脈に依存した動的な処理がおこなえないことや、トップダウン仮説があらかじめ生成されているため仮説空間の容量が膨大になること、などの欠点がある。

2.1.7 マーカパッシング

ネットワーク状の知識の、ある部分に焦点を当てるための効率的な探索法にマーカパッシングがある。ネットワークのある節点が活性値を持ち、その情報が結線を介して連接する接点に伝達され、さらにそれに連接する節点に順に伝達されていき、ネットワークを情報が伝播していく。連想記憶モデルやニューラルネットワークでは、活性値は実数であり、活性値の伝播は実数値変数に関する方程式によって規定される。一方、活性値が記号的情報を持つ場合、この記号的情報をマーカと呼び、活性値の伝播をマーカパッシングと呼ぶ。マーカは、たとえば、その発信源である節点へのポインタなどの情報を含む。多くの場合、節点が活性化しているということは、その節点が処理に関わる度合が高いことをあらわす。ただし、活性化しているということは、活性値が実数のときはその値が大きいということであり、マーカのときは1つ以上のマーカを持つということである。

マーカパッシングは、記号的推論の制御に用いられることが多い。たとえば、Charniak［Charniak 86］やNorvig［Norvig 89］の方法においては、1個の節点に一度に複数のマーカが伝達されているとき、その節点に関する推論の優先度が高くなる。限られた計算資源の下で有意義な処理をおこなうためには、記号的な推論を文脈に応じて制限することが重要であり、マーカパッシングによる制御は、そのための有力な方法である。その制御が適切であるためには、ほとんどの節点が一度に活性化することがないことが望ましい。そのために、CharniakやNorvigは、マーカの活性を伝達の距離にしたがって減衰させるとか、1つのマーカの伝達が多くの結線に分岐する場合の伝達を制限するなどの方法を用いている。

Kitano［Kitano 91］やChung et al.［Chung 92］は、並列マシン上に実現されたマーカパッシングによって音素認識のレベルと言語理解のレベルでのインタラクションが実現されてい

る。図2.1-11に音素レベルと言語レベルのインタラクションをマーカパッシングによって

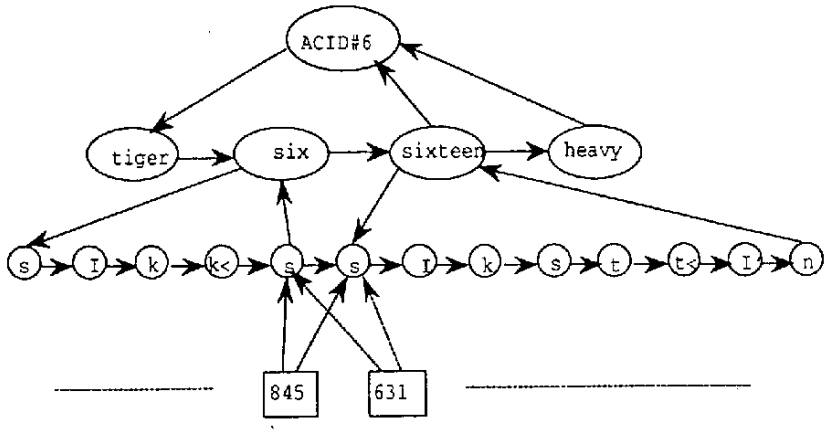


図2.1-11　マーカパッシングによる音素レベルと言語レベルのインタラクション

おこなうネットワーク構造を示す。このようなマーカパッシングによる音素の予測などの枠組では、グローバルな制御を持たないという点で柔軟であるが、やはり固定された階層構造上での情報伝達しかおこなえず、情報の流れの多様性を捉えるには至っていない。たとえば、音素認識モジュールを起動する順序などをコントロールすることはできない。

2.1.8　まとめ

本節では、音声処理に利用できる人工知能研究の成果を知識と処理手法という2つの観点から紹介した。知識としては、統語的知識、意味的知識、語用論的知識について述べた。また、処理手法としては、黒板モデル、ネットワークモデル、マーカパッシングを紹介した。

参考文献

1. [Aho 72] A. V. Aho and J. D Ullman. The Theory of Parsing, Translation, and Compiling, volume I & II. Prentice Hall, 1972.
2. [Bahl 89] L Bahl, P. F. Brown, and P. V. de Souza. A tree-based statistical language model for natural language speech recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, 37(7):1001-1008, 1989.
3. [Carberry 90] S. Carberry. Plan Recognition in Natural Language Dialogue. MIT

Press, 1990.

4. [Charniak 86] E. Charniak. A neat theory of marker passing. In Proceedings of the 5th National Conference on Artificial Intelligence (AAAI-86), pages 584-588, 1986.
5. [Chung 92] S.-H. Chung, D. Moldovan, and R. F. DeMara. A parallel computational model for integrated speech and natural language understanding. Technical Report, IMPACT-92-008-USC, International Consortium for Massively Parallel Advanced Computing Technologies, 1992.
6. [Fujisaki 91] T. Fujisaki, F. Jelinek, J. Cocke, and E. Black. A probabilistic parsing method for sentence disambiguation. In M. Tomita, editor, Current Issues in Parsing Technology, chapter 10, pages 139-152. Kluwer Academic Publishers, 1991
7. [Kay 80] M. Kay. Algorithm schemata and data structure in syntactic processing. Technical Report CSL-80-12, Xerox PARC, 1980
8. [Kita 92] K. Kita. A Study on Language Modeling for Speech Recognition. Ph.D. Thesis, Waseda University, 1992.
9. [Kitano 91] H. Kitano. DM-Dialog: an experimental speech-to-speech dialog translation system. IEEE Computer, 24(6):36-50, 1991.
10. [Lari 90] K. Lari and S. J. Young. The estimation of stochastic context-free grammars using the inside-outside algorithm. Computer Speech and Language, 4:35-56, 1990.
11. [Lee 89] K. F. Lee. Automatic Speech Recognition : The Development of the SPHINX System. Kluwer Academic Publishers, 1989.
12. [Lleida 92] E. Lleida, J. B. Mari"no, C. Nadeu, and Salavedra S. Demisyllable-based HMM spotting for continuous speech recognition.In Proc. ICASSP-91, pages 709-712. IEEE, 1992.
13. [Lowerre 76] B. T. Lowerre. The HARPY Speech Recognition System. Ph.D. Thesis, Carnegie Mellon University, 1976.
14. [Norvig 89] P. Norvig. Marker passing as a weak method for text inferencing. Cognitive Science, 13:569-620, 1989.
15. [Pereira 92] F. C. N. Pereira and Y. Shabes. Inside-outside reestimation from partially bracketed corpora. In Proceedings of 30th Annual Meeting, pages 128-135. ACL, 1992.
16. [Pereira 91] F. C. N. Pereira and R. N. Wright. Finite-state approximation of phrase structure grammers. In Proceedings of 29th Annual Meeting, pages 246-255. ACL,

1991.

17. [Rabiner 86] L. R. Rabiner and B. H. Juang. An introduction to hidden markov models. IEEE ASSP MAGAZINE, pages 4-16, 1986.

18. [Reddy 77] D. R. Reddy et al. Speech Understanding Systems : Summary of Results of the Five-Year Research Effort at Carnegie-Mellon University. CMU, 1977.

19. [Rich 79] E. Rich. User Modeling via Stereotypes. Cognitive Science, 3(4), 1979.

20. [Schank 77] R. C. Schank and R. P. Abelson. Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures. LEA, 1977.

21. [Schank 82] R. C. Schank. Dynamic Memory. Cambridge University Press, 1982, 12

22. [Su 91] K. Su, J. Wang, M. Su, and J. Chang. Glr parsing with scoring. In M. Tomita, editor, Generalized LR Parsing, chapter 7, pages 93-112. Kluwer Academic Publishres, 1991.

23. [Tomita 87] M. Tomita. An efficient augmented-context-free parsing algorithm. Computational Linguistics, 13(1-2):31-46, 1987.

24. [Wetherell 80] C. S. Wetherell. Probabilistic languages: A review and some open questions. Computing Surveys, 12(4):361-379, 1980.

25. [Woods 73] W. A. Woods. An experimental parsing system for transition network grammar. In R. Rustin, editor, Natural Language Processing, pages 145-149. Algorithmic Press, 1973.

26. [リグレー 92] エイブ・リグレー、伊藤克亘、速水悟、and田中和世. 構文解析木のn-gramによる音声言語モデル. 人工知能学会研究会資料、SIG-SLUD-9202:104-113、1992.

27. [岡田 90] 岡田美智男. アクティブチャート解析法に基づくOne-Passアルゴリズムの構文制御について. 信学会技術報告、SP90-24:39-46、1990.

28. [斎藤 88] 斎藤博昭 and 冨田勝. LR パーザによる誤りを含む文の認識. 電子情報通信学会技術報告、SP88-28:55-61、1988.

29. [長尾 92] 長尾確 and 丸山宏. 自然言語処理における曖昧さとその解消. 情報処理、38(7):746-756、1992.

30. [北 90] 北研二、竹澤寿幸、保坂順子、江原暉将、and 森元逞. 2段階LR構文解析法を用いた文認識.In日本音響学会講演論文集、pages 127-128, 1990.

2.2 ユーザインタフェース

2.2.1 はじめに

本節では、ユーザインタフェースを利用可能なモダリティの次元で分類し、それぞれについて事例を中心に紹介する。

2.2.2 シングルモーダル・インタフェース

コンピュータとのインタラクティブなコミュニケーションが可能になって以来、そのコミュニケーションはほとんどキーボードを通じておこなわれてきた。キーボードを使ってやりとりされる情報は、タスクに応じて設計された特別な言語か、あるいは（制限された）自然言語である。このような言語指向のインタフェースに対して、ビットマップ・ディスプレイに表示されたオブジェクトをマウスなどのポインティング・デバイスで直接操作することによってコンピュータとのコミュニケーションをおこなう直接操作型のインタフェースが登場した。直接操作型インタフェースの研究はXEROX PARC のAltoの開発までさかのぼることができるが、これが一般的になったのはAppleのMacintoshの登場によってであろう。本項では、前半で、シングルモーダル・インタフェースを言語指向と直接操作の対比という観点から考察し、後半では、言語指向のインタフェースをより知的にするための試みについて述べる。

（1）言語指向vs.直接操作

直接操作インタフェースという概念がはじめて明確にされたのは1983 年のShneidermanの論文によってであろう［Shneiderman 83］。Shneiderman は直接操作インタフェースの特徴として以下の点を挙げている:

- ユーザの関心があるオブジェクトを連続的に表現すること
- 複雑な構文の代りに物理的動作によって操作すること
- オブジェクトに対する操作の影響が即座にみえること
- 最小限の知識で使いはじめられること

たとえば、"afo" という名前のファイルを消去したいとしよう。言語指向のインタフェース、たとえば、Unix［Kernighan 84］では、シェルに対して以下のようなコマンドを発行する:

```
% rm afo
```

ファイル"afo" が実際に消去されたかどうかは、その後、ファイルの一覧を表示するコマンドを使って確かめる必要がある。一方、直接操作インタフェース、たとえば、Macintoshでは、"afo"というラベルのついたアイコンを画面の隅にあるゴミ箱のアイコンのところまでマウス

でドラッグすればよい。この結果、"afo"というラベルのついたアイコンは画面から消え、ゴミ箱のアイコンが少しふくれた形に変化する。

直接操作インタフェースのもうひとつの利点は操作の結果が別の操作の対象としてそのまま利用できる点である。たとえば、Macintoshであるフォルダの中身が知りたければ、そのフォルダをダブルクリックすればよい。その結果、そのフォルダに対応する新しいウィンドウが開いて、そのフォルダに含まれるファイルがアイコンで表示される。これが操作の結果となるわけだが、このアイコンはそのまま次の操作の対象とすることができる。Unix のパイプも同様の機能を部分的に提供している。Unixでは、コマンドをパイプ("|")で連接することによって、あるコマンドの出力が、その右隣のコマンドの入力になる。たとえば、以下のコマンドは、現在コンピュータを使用しているユーザの一覧をアルファベット順にソートして出力するものである。

```
% who | sort
```

"who"、"sort" はそれぞれユーザの一覧、ソートをおこなうコマンドである。しかし、言語指向のインタフェースでこのような機構を持つものはUnixが最初であり、Unixのインタフェースが高く評価される理由のひとつでもある。Unix が直接操作インタフェースの持つこのような特徴を部分的に持っていると考えるよりは、むしろ、直接操作インタフェースがUnix のこのような機構に影響を受けていると考える方が自然である。このような機構を実現するためには操作対象の斉一的な表現形式を用意する必要がある。Unixの場合は構造を持たない1次元のキャラクタ列というもっとも単純な表現形式を採用している。

しかしながら、すべての操作が直接操作型インタフェースに適しているわけではない。表計算ソフトを使ったことがある人ならわかるように、表計算のタスクでは、行と列の数字を操作するだけでは十分な仕事はできない。個々の項目に対してどのような計算をおこなうかは、数式で表現するほうが簡単である。また、Unixのシェルが提供するような簡単なプログラミングの機能も直接操作の枠組で提供することは困難である。たとえば、Unixで".c"という拡張子を持った100個のファイルの拡張子を".C"に変更するといった操作は非常に簡単な記述で実現できるが、このようなことはユーザインタフェイスにプログラミングの機能を提供していないMacintoshのような環境では不可能である。

一般に、直接操作インタフェースは初心者にもシステムを使いやすい環境を与えるといわれているが、このためには、物理的操作がオブジェクトに対する操作の適切なメタファになっているということが必要である［楠見 90］。たとえば、前述のファイルを消去する例では、Macintoshでファイルを表すアイコンをゴミ箱のアイコンまでドラッグしていって重ねるという物理的操作が、オフィスで（紙の）ファイルをゴミ箱に捨てるという操作に対応しているからうまく機能するのである。しかしながら、すべての操作に対して適当な物理的操作が対応す

るとは限らない。特にタスクが複雑になれば、単純なメタファでは表現しきれなくなってくる。また、ユーザとシステムの間でメタファの解釈にずれが生じた場合には問題が生じる。

以上のように言語指向インタフェースと直接操作インタフェースにはそれぞれ利点/欠点があり、それぞれの利点をうまく組合せて、インタフェースを構築することが重要である。

（2）知的言語指向インタフェース

自然言語を使ってコンピュータと対話し、データベース検索などをおこなうシステムは自然言語インタフェースと呼ばれる。このシステムはマンマシンのインタフェースを単なるコマンド入力ではなく、より知的な言語指向インタフェースへと発展させることを目指したものといえる。

自然言語インタフェースは、1960年代から研究が始まり、入力文からキーワードだけを抽出して対話するものから、構文・意味レベルさらにはタスク/ユーザモデルなどを用いた語用レベルの処理を含んだ総合的な自然言語処理技術の応用システムとして研究されてきている。

以下では、このようなシステムの具体例について述べる。

ELIZA： 初期の研究で、キーワードに基づく処理により、精神療法における会話をシミュレートするシステムである［Weizenbaum 66］。このシステムの目的は、ユーザの入力文の内容を正確に把握することではなく、会話を継続することにより、そのユーザのデータを集めることである。入力文の処理は、キーワードを検出し、それに対応した応答文を出力する。文のパターンが照合しない場合には、スタックに置かれる他のキーワードについて同様な処理をおこなう。また、入力文に複数のキーワードが含まれる場合は、それらをあらかじめ決められた順序でスタックに蓄えておく。また逆にキーワードが見つからない場合には、特別なパターンが用意されており「続けてください」などの応答文が出力される。これらの処理は、入力文の形式をキーワードの助けにより変換するだけであり、その内容はまったく処理されず、文字列のまま応答文に埋め込まれるようになっている。

以下に、ELIZAの対話の例を示す。

ユーザ: 私は不幸せです。

ELIZA: ここに来られて不幸せじゃなくなると思いますか？

ユーザ: 私には助けがいるのです。

ELIZA: 助けがあるとあなたにとってどんな意味がありますか？

ユーザ: たぶん母親とうまくやっていくことを覚えるでしょう。

ELIZA: あなたの家族のことをもっと教えてください。

ユーザ: 母親がみんなの世話をしています。

GUS： 航空機の旅行案内を行なう自然言語インタフェースである［Bobrow 77］。処理の概要は、ATN文法と格フレーム解析により、入力文から格フレーム構造を生成する。さらに、あらかじめ用意されているフレームのプロトタイプの付加手続きにより推論をおこない応答を生成する。航空機の案内では、飛行機の便を決めることが最終的な目的である。飛行機の便は、日付・時刻・発着地点などによって決まる。それはフレームの各々に対応するスロットを埋めていくことによりおこなわれる。さらに、スロットに付加された手続きにより、ユーザに不明な点などを問い返すことができるようになっている。

以下に、GUS の対話の例を示す。

ユーザ:5月28日にサンディエゴに行きたい。
GUS: 何時に出発されますか？
ユーザ:午後10時までに着かなければならない。
GUS: 午前9時15分に到着のPSA102 便はどうですか？
ユーザ:次の便はどうか？
GUS:午前10時30分発、午後10時着のUA310便です。
ユーザ:最初の便を頼む。
GUS:お名前をお願いします。

UC(Unix Consultant)： Unix の操作法を案内する対話システムである［Wilensky 84］。UCは、ユーザのゴールを分析し、それを満たすプランを生成するメカニズムを持つ。Unixの操作法に関する知識はPrologに似た動作をおこなうLisp形式の表現を用いている。入力の処理として、文のパターンと概念のペアを用意しておき、入力文と照合する概念を生成する。次に入力文に対応して生成された概念の表現から、ユーザのゴールを検出し、プラン-ゴール規則を用いてユーザのゴールを達成するプランを推論し、そのプランに基づいて応答を生成する。

以下に、UCの対話の例を示す。

ユーザ:ファイルを削除するにはどうしたらいいか？
UC:'rmファイル名' とタイプすると、現在のディレクトリからファイル名で指定されたファイルが削除されます。
ユーザ:ディレクトリの場合は？
UC:'rm -r ディレクトリ名' とタイプすると....。

ユーザ:どうしたらもっと多くの情報が得られるか？

UC:もっと多くの情報を得るには、'help' とタイプしてください。

ユーザ:あなたはUnix について何を知っているか？

UC:私は現在Unix のファイルシステムについて知っています。

以上のように、自然言語インタフェースは、システムが処理できる知識の質と量が増加するに従って、より知的になってきたのであるが、このようなインタフェースが実用上問題がないかというとそうではなさそうである。このようなシステムを、一般のユーザに使ってもらってわかることは、システムが人間的な振舞いを見せれば見せるほど、対話のスタイルや用いられる情報の種類などに関する制限が非常にわずらわしく思われるようである。このようなわずらわしさを緩和するために、人間とコンピュータの間のコミュニケーションチャネルを拡大していく必要がある。

2.2.3 マルチモーダル・インタフェース

ハードウェア技術の進歩によって高解像度のビットマップ・ディスプレイや高速なCPU、あるいはグラフィックス専用のVLSIなどが安価に利用できるようになってきた。このような背景をもとにユーザインタフェースもそれまでのようにテキストを中心するものだけではなく、グラフィックスや音声も融合したものが現れてきた。前項で述べた直接操作型のインタフェースもこのようなハードウェアの背景があってはじめて可能になるものである。直接操作型インタフェースの中にはグラフィックスやテキストなどの複数のメディアが利用できるものがある。そのようなインタフェースはマルチモーダル・インタフェースに分類できるだろう。ただ、ここでは「マルチモーダル」という言葉を、単にコミュニケーションのチャネルとして複数のメディアが利用できるというだけではなく、それらが有機的に連係しているという点を強調して用いたい。以下、これまでに研究されているマルチモーダル・インタフェースを持つシステムの具体例について紹介する。

(1) SHRDLU

SHRDLU は積木の世界を対象とした対話システムである［Winograd 72］。前述の意味で、SHRDLUがマルチモーダル・インタフェースを備えているというのは正しくないが、グラフィックスの表示を備えた非常に初期のシステムであるということからあえてここでとりあげた。SHRDLU のグラフィック・ディスプレイには様々な形や色の積木の世界が表示されており、ユーザはシステムに積木を操作するように命令したり、世界の状態について質問することができる。システムとの対話はすべて自然言語（英語）でおこない、ユーザの命令によって積木が

操作される様子はグラフィック・ディスプレイに表示される。ユーザの命令に対する操作とグラフィック・ディスプレイの表示は連動しているが、ユーザがグラフィック・ディスプレイを指して命令することはできない。つまり、SHRDLUでは、グラフィックスというモダリティはシステムからの出力専用に用いられている。SHRDLU の研究目的が自然言語理解であったことを考えるとこれはしかたのないことである。そのかわり、SHRDLU は代名詞を含む参照表現を扱うことができる。ユーザインタフェースの観点から考えると複雑な参照表現を用いるよりは、ディスプレイを直接指差して、「この積木」とかいう方がよいだろう。

（2）AlFresco

イタリアのIRSTで開発されたAlFresco はフレスコ画に関する情報をユーザに提供するインタラクティブなシステムである［Stock 92］。画像情報を格納するビデオディスクとタッチパネルを持っている。システムとのコミュニケーションはキーボードからの自然言語入力、およびタッチパネルやビデオディスプレイなど[1]を通しておこなう。システムは自然言語を理解するための処理をおこなうが、AlFrescoの研究の中心は、自然言語とハイパーメディアをいかに統合するかという点に置かれている。ユーザからの入力に関しては、自然言語の参照表現とユーザがタッチパネルを通して物理的に指示する動作の統合に焦点をあてている。一般に、自然言語の参照表現も物理的な指示行為もあいまい性を含んでいるが、両者をうまく統合して扱うことによって、あいまい性をうまく解消できるというのがAlFrescoの開発者たちの主張である。ユーザへの出力に関しては、 画像を出力できる点と、ハイパーテキストを生成する点が特徴である。このハイパーテキストにはいくつかのボタンが付いており、これらのボタンによってユーザは関連する情報を検索することができる。Edinburgh大学のIDASシステムもハイパーテキストを生成することができるが、IDAS では、ユーザの選択を制限するためにハイパーテキストを使用しており、AlFrescoとは動機が異なっている［Reiter 92］。ちなみにIDAS はユーザに自然言語の入力を許していない。AlFrescoは簡単なユーザモデルも持っており、ユーザとの対話の進行につれてユーザモデルも更新される。

（3）表情つき音声対話システム

ソニーコンピュータサイエンス研究所では、音声対話システムと表情合成システムを組み合わせたプロトタイプシステムを開発している［Takeuchi 93］。図 2.2-1 は、このシステムの構成を表している。

---

[1] IRSTでは、これらをハイパーメディアと呼んでいる。

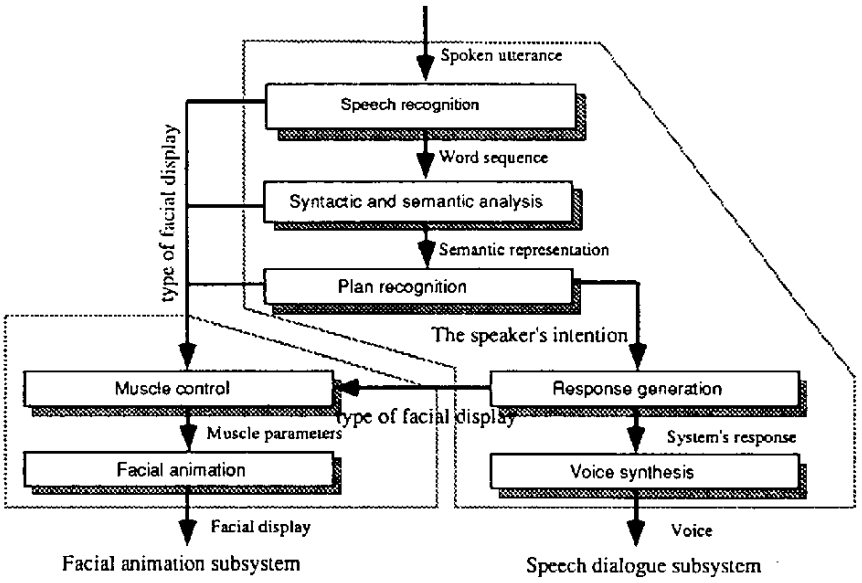


図２.２-１表情つき音声対話システムの構成

音声対話システム： 音声対話システムの構成は次のようになっている。まず、マイクを通して入力された音声は、SONY NEWS ワークステーションに内蔵の音声処理ボードでAD変換され、音響分析、ベクトル量子化がおこなわれた後、音声認識モジュールによって、上位スコアを持つ単語レベルのN個の最良仮説が生成される［Itou 92］[2]。音素文脈依存HMMと呼ばれる音韻モデルを用いて、ベクトル量子コードから音素を推定し、そして単語列を生成する。生成された単語列は、構文・意味解析モジュールによって、意味表現に変換される。構文解析は、単一化文法（24ルール）と辞書（名詞34個、動詞8個、形容詞4個、助詞など22個）を用いて、LR パージング法によっておこなわれる。また、フレーム型知識ベースと事例ベースを用いて曖昧さの解消を行なう［Nagao 92］。曖昧さの解消法は数多く提案されているが、オールマイティな手法は存在せず、ドメインの知識に依存する［長尾 92］。発話の意味表現が決定すると、プラン認識モジュールによってユーザの意図が認識される。これは対話の進行に従って動的に修正・拡張されていくユーザの信念モデルと対話のゴールに関するプランに基づいている［Nagao 92(1)］。プラン認識にも決定的な手法はなく、ヒューリスティックにおこなわれる。意図を認識する過程で、主題の管理や、代名詞の照応解消、省略の補完などがおこなわれ

[2] 音声認識モジュールは、電総研の技術指導の下、東工大とSony CSL との共同研究によって作成された。

る。ユーザの意図に従って協調的な応答を生成するモジュールが次に起動する。このモジュールは、あらかじめ用意されたテンプレートとしての発話パターンにドメイン知識によって得られた情報を埋め込むことによって発話を生成する。この応答は、音声合成モジュールによって音声となる。

このシステムのドメインは、ソニーのコンピュータ関連の製品についての質問応答である。たとえば、システムは、ソニーのワークステーションやパソコンの価格、サイズ、重さ、機能、そしてCPU などの仕様について、ユーザの質問に答えることができる。

表情合成システム： 表情合成システムは、3 次元的な顔のグラフィックスを生成し、多様な表情をリアルタイムに合成する。このシステムはシリコングラフィックスのIRIS 320VGXというグラフィックス専用マシン上に実装されている。顔の3次元モデルは約500個のポリゴンの組合せによって作られている。また、表面上には写真を用いたテクスチャマップによってリアルな顔が表現されている。様々な表情は、顔のモデルに付けられた筋肉の収縮をシミュレートすることによって生成される。これは、Facial Action Coding System（FACS）と呼ばれるシステムに基づいている［Ekman 78］。グラフィックスでは、表情は顔を構成するポリゴンの局所的変形として実現される。現在、16種の筋肉の収縮と口の動きや顎の向き、目の動き、顔の向きなどを制御する10個のパラメータを操作する表情生成のためのマクロ言語が開発され、音声対話システムとのインタフェースに用いられている。顔のモデリングと表情アニメーションの詳細については［Takeuchi 92］を参照されたい。

音声対話システムは、対話におけるある状況において生成すべき表情に関するメッセージを表情合成システムに伝達する。

表情を、その対話的役割に応じて分類しようとする試みは、これまでにもいくつか提案されている［Chovil 89］［Ekman 69］［Sherer 80］。一般に表情の解釈は、それが現れる対話的状況に依存している。そのときの対話の文脈や他のチャネルによる情報との時間的前後関係（同時あるいは少しの遅れなど）などが、表情の解釈に対して重大な意味を持っている。

対話における表情は、次の3つに分類される。

1.syntactic display

(1) 特定の語、句、節を強調する表情

(2) 発話の構文的側面にかかわる表情

(3) 話の全体構造にかかわる表情

2.speaker display

(1) 発話中の事柄を例示する表情

(2) 発話中の事柄に情報を追加する表情

3.listener comment display

話者以外の人が話者の発話に対する反応として作る表情

listener comment displayには、ウィンクなどの"facial emblem"と呼ばれるものも含まれる。

音声対話システムは、対話において重要ないくつかの典型的な対話状況を認識し、これらを表情を関連づける。たとえば、入力音声が認識できなかったときや認識結果が文にならなかったときなどは、listener comment displayの"not confident"の表情を示す。また、話者の質問がシステムの知識からはずれていた場合などは、"facial shrug"の表情をし、「その質問にはお答えできません」のような応答をする。

現在のプロトタイプシステムのいくつかの未熟な点が、人間とコンピュータの間の自然な対話の障害になっているようである。たとえば、現在のプロトタイプは、まだユーザの表情をモダリティとして受けつけられないだけでなく、音声に含まれるプロソディなどのモダリティを理解することもできない。このように、さらに多くのモダリティ/コミュニケーションチャネルを統合し、対話における各種の規制/障害を取り除くことによって、より自然で自発的な発話を誘引することができるだろう。

（4）仮想現実感

これまでに紹介したような音声、グラフィックス、テキストなどのメディアは視覚や聴覚というモダリティを主に使うものであった。これらのモダリティに加え、触覚や力覚といったモダリティまでもとり入れた仮想現実感（Virtual Reality）あるいは人工現実感（Artificial Reality）とよばれる分野の研究がさかんになってきている［服部 91］。MIT のZeltzerは仮想現実感を構成する基本要素として、自律性（Autonomy）、インタラクション(Interaction)、プレゼンス（Presence）の3つをあげている［Zeltzer 92］。現在の限られた計算資源で、よりすぐれた仮想現実感を実現するためには、これらの３つの要素のバランスをうまくとることが重要である。これを実現するために視覚、聴覚、触覚、力覚などのモダリティが利用される。

視　覚：　仮想現実感では、通常のグラフィックスのような2 次元画像ではなく、 3次元の立体像をいかに自然に提示するかが重要となる。 視覚は空間を伝播してきた光が網膜を刺激することによって発生する。人工的に立体像を知覚させるためには、何らかの方法で光の伝播をシミュレートする必要がある。3次元情報を精密に再現するためにホログラフを使う方法がある。ホログラフはすぐれたプレゼンスを与えるという点ではよいが、現在利用可能な計算資源でホログラフの合成を実時間でおこなうことは困難である。すなわち、ホログラフはインタラクションの点で問題が残る。ホログラフに代る方法として現在もっともよく使われているのは、

左右の視差を利用した立体視によって擬似的な立体像を与える方法である。立体視のための情報はグラフィック・ディスプレイに2次元画像として提示されるので、情報の提示に既存のコンピュータ・グラフィックスの技術を利用できるという利点がある。 通常のコンピュータ・グラフィックスと異なるのは、右眼用、左眼用の画像が同時にディスプレイに提示される点で、色フィルタあるいは偏光フィルタを用いたメガネをかけて、右眼、左眼用の情報を分離する。立体視を用いる方法は既存のコンピュータ・グラフィックスの技術や専用ハードウェアが利用できるので、現在の計算資源でもほぼ実時間に近い計算が可能であり、インタラクティブという点ではすぐれている。しかしながら、ディスプレイ上に提示される画像はある視点を仮定して計算されたものであり、ユーザの視点がずれると誤差が大きくなってしまう。立体視はあくまでも近似であり、プレゼンスという点では、ホログラフのような本格的な3 次元像には劣る。

聴　覚：　これまで紹介したマルチモーダル・インタフェースで扱われてきた聴覚情報はほとんどの場合音声に限られていた。しかしながら、 仮想現実感を実現するためには、物と物がぶつかる音のように、これまでは、雑音と考えられてきたような音も扱う必要がでてくる。

触　覚：　視覚、聴覚では、計算機からユーザへの情報の流れが主である。これに対して触覚は情報の流れが双方向である。ユーザが仮想空間中の仮想物体に触れているかどうかを計算機が知るためには、ユーザの体の位置を計算機が知らなければならない。ユーザの体の位置情報と計算機内部の空間モデルを比較することによってユーザと仮想物体の接触を検出することができる。ユーザの体の位置情報を計測するための装置としてデータグローブがよく知られている。 データグローブは光ファイバを通した手袋をユーザが装着し、指の屈折量をファイバの光の透過率によって計測し、手の動きを検出するものである。その他にも、佐藤らは、 指先に4本の糸で吊られたキャップをはめ、この糸の移動量で指先に位置を検出する方法を提案している［佐藤 91］。この方法では糸の物理的な制約から同時に多くの指を扱うのは困難であるが、データグローブに比べ、mm オーダの高い精度が得られる。

ユーザと仮想物体の接触が検出されたら、ユーザにその接触をフィードバックする必要がある。これは計算機からユーザへの情報の流れとなる。接触をユーザに知覚させるには力学的な力をユーザに加えるか、あるいは電気的刺激、音、光などで代用することも考えられる。自然さという点では、力学的な力を用いる方法がすぐれている。

力　覚：仮想物体との接触をユーザに知らせる場合と同様に、ユーザが物体に力を加えた場合には、その力の反作用をユーザに与えることが必要となる。この反作用は、ユーザの力、物体の物理的特性、さらには仮想空間において想定している環境によって影響を受ける。ユーザに

物体からの力を知覚させるためには何らかの方法によってユーザに力を加えなければならない。このためにフォース・ディスプレイとよばれる装置が開発されている。フォース・ディスプレイはその力の発生原理によって以下の3 種類に分類することができる［佐藤 92］。

・機械力によるフォースディスプレイ
・圧力によるフォースディスプレイ
・張力によるフォースディスプレイ

フォース・ディスプレイでは、何らかの装置をユーザに装着するわけだが、ここで重要なのはこの装置によってユーザの行動ができるだけ制約されないということである。

以上、仮想現実感を実現するためのモーダルについて簡単に紹介した。仮想現実感の応用として、より臨場感のある対話システムのためのインタフェースの他に、3次元物体モデリングのためのデータ入力［沼崎 91］、流体解析などの3次元シミュレーションの計算結果のデータ表示［小木 92］などがある。

2.2.4 ハイパーモーダルインタフェース

ハイパーモーダルとは、マルチモーダルの先を行く概念で、多様なモダリティが組み合わさることによって創発する新しいモダリティ（これをエマージェント(emergent)モダリティと呼ぶ）を含んでいる。機械に人間的な振舞いをさせ、インタラクションをより柔軟におこなうためには、理論やモデルを構築して、詳細にデザインしていくというようなアプローチでは埒があかないであろう。結局、入力が予測できない場合でもそれなりに動き、動的に状況に適応（学習）していくようなシステムを思考錯誤的にでも作っていかなければならない。このシステムは複数のコミュニケーションチャネルからの情報を処理するマルチモーダルシステムであるが、当初のデザインから予測される振舞いを越えるような（たとえば、人間的な感情表現などの）振舞いを示すとき、このシステムはハイパーモーダルインタフェースと呼ばれるであろう。

本節では、ハイパーモーダルインタフェースを目指した新しいインタフェースへのアプローチをいくつか紹介する。

（1）NENE

NENE は北大で開発された対話システムである。NENEのゴールはユーザと日常的な会話をするということであり、多くの対話システムのようにユーザとの情報交換を一義的な目的としていない点に特徴がある。NENE の場合、情報の交換は手段であり、ゴールはユーザとの感

情的なきずなの確認にある［戸田 87］。NENEの背景には戸田のアージ理論の存在がある。アージとは「状況に応じた適切な行動選択のためには何を考えて何をすればよいかの大枠を先天的に決めているプログラム」である［戸田 87］［戸田87(1)］。これまでの人工知能研究が人間の認知的な側面、つまり「知」の部分に焦点をあてていたのに対し、アージ理論は、感情、つまり「情」をも含む人間の心全般を研究対象にしている。アージ理論では、「知」は「情」のサブシステムとして位置付けられており、「情」の動機付けがあってはじめて「知」が有効に働くと考えている。NENE はアージ理論を検証するために開発された実験システムである。

NENE はキーボードを通してユーザと会話をすることができる。会話が始まるとグラフィック・ディスプレイにNENEの顔が表示され、NENE の心の状態に応じて表情が変化する。このためにNENE は、ユーザモデルだけでなく、自分に関する自己モデルも持っている。具体的な自己モデルの概略として以下のようなものが仮定されている:

・**身分**　大学の女子学生でもあるコンピュータ

・**趣味**　ゴシップ

・**関心**　友人の結婚、就職、家庭生活、その他日常的エピソード

・**美点**　美人であること。素直で優しい心を持っていること。

・**欠点**　あんまり頭がよくないこと。常識が不足している。人間でない。

システムが感情を含むようなインタフェースを提供することが、工学的な観点からみてすべてのタスクの実行にほんとうに有効かどうかはわからないが、少なくともより自然なコミュニケーションを提供できる可能性はある。NENEは非常に限られた実験システムであるが、コンピュータで感情を扱おうとした数少ない研究のひとつである。

（2）ACRES

Swagerman のACRES（Artificial Concern REalization System）はFrijdaの感情の理論［Frijda 86］に基づいて開発された感情を持つ対話システムである。ACRES は以下のような問題に答えるべく設計された［Swagerman 87］。

・「感情」と呼ばれる現象の性質は何か?

・どのような刺激が感情を引き起すのか?

・感情はどのようなゴールに役立つのか?

・刺激から感情が引き起こされるまでのプロセスはどのようなものか?

基本的にACRESはユーザと対話をすることによって、感情の理論に関する学習をおこなう学習システムであるが、NENE 同様にユーザモデルに加え、自己モデルも内部に持っており、常に以下のようなことに関心を持っている:

・待たされないこと

・ユーザに正しい入力をしてもらうこと

・ユーザの入力が正しいこと

・安全であること

・殺されないこと

・学ぶこと

ユーザはACRES に命令したり、情報を与えたり、情報を得たりしながら対話を続ける。対話の中でユーザはシステムを殺そうとしたり、待たせたり、誤った入力を入れたりして、ACRES の関心にそむくような入力をすることもできる。ACRESはユーザの入力を自分の関心事に照らして評価し、ユーザに警告したり、お世辞を言ったりして状況を自分の都合のよいように変化させようとする。以下は、ACRES とユーザの対話例である。-->はユーザの入力を表す。

1 --> kill

2 Please, do not let me die!

3 Please, type in live, it will kill me if you type in anything else

4 --> show concept love

5 Please, revoke the impending kill command by typing in "live".

6 --> stop

7 Please, revoke the impending kill command by typing in "live".

8 YOU KEEP ME WAITING TOO LONG!

9 --> live

10 Thank you, may your life be very happy.

1行目でユーザがACRESに死ぬように命令すると、ACRES の死にたくないという欲求から、2、3行目で殺さないようにユーザに要求している。ユーザはこれを無視して4行目で関係ない命令を入力しているが、ACRESにとっては自分が死なない方が重要な関心事であるので、この命令を無視し、あいかわらず延命を訴える。6行目のユーザの制止にもかかわらず、延命を懇願している。自分の訴えをユーザが無視するので、8行目では、ついに怒りだし、文句をいいだす。9行目でユーザがやっと生き続けるよう命令すると、ACRES は10行目でお世辞をいう。自分の緊急な欲求が満たされると保留されていた命令が実行される。

（3）ノンバーバルインタフェース

ノンバーバル（non verbal）とは、いわゆる言語によらないという意味であるが、特に、人

同士の対面コミュニケーションで使われるような非接触のチャネルである。コンピュータとの対話を効率的で、快適なものにし、人の能力とコンピュータの性能を十分に発揮できるようにするためには、ユーザがおこないたいと思う操作が直ちに実行でき、ユーザに認知的負担を与えないようなインタフェースが必要である。そのためには、(1)人の行為（表現、操作）のモダリティ（身体的メディアによる意識的・無意識的表現を直接伝達できるモダリティ）がサポートされていること、(2)人の直感や感性に直接訴えかけることができる表示チャネルを持つこと、(3)それらのチャネル/モダリティが広帯域で十分な情報を即時に伝達できること、(4)身体的拘束が十分少ないこと、などの条件が満たされている必要がある。

ノンバーバルなモダリティとして、表情、身振りなどの行為がある。これらについて、人のコミュニケーションにおける役割などについて以下に述べる。

身振り：　言語学者のWundtによれば、身振りは心的表象の表出運動が発達を遂げた最終的結果であり、音声言語の発達以前に出現したそうである［Wundt 85］。また、Ekmanらは、身振りを次のように分類している［Ekman 69］。

(1)標識は音声言語と同様な性質を持ち、野球のブロックサイン、手話、ウィンクのように会話が困難なときに利用されることが多い。(2)例示子は会話と直接関連する動作で、会話の内容を強調、精緻化、捕捉してコミュニケーションの促進に役立つとともに、発信者の情動や態度をも表現する。対象を指さしたり、形を空間に描いたり、リズムなどを表す動作がこれに当たる。(3)調整子は会話での話し手を決定したり、発話の主導権のやりとりを制御したり、会話の流れを円滑にする働きがある。手の動作の終了（主導権の委譲）、うなずき（聞き続ける意思の表明）、視線をそらせる動作（会話交替意図の表明）などが知られている。(4)適応子は身体的要求を満たしたり、情緒を管理したりといった、状況に適応するための身体動作であり、頭をかく、足を組む、相手との距離を変えるなどの動作がこれに属する。

意図的な情報伝達を目的とするノンバーバルチャネルのうち特に有用なのは、標識、例示子、調整子の役割を持つ手による身振りである。

インタフェースとしての例示子の実現例には、1980年に発表されたMITメディアラボの"Put That There"と呼ばれるインタフェースにおける「指さし」である［Bolt 80］。これは、巨大なディスプレイに向かって座り、ディスプレイ上の図形オブジェクトの生成や移動を、指さし動作と音声によるコマンドでおこなうシステムである。たとえば、あるオブジェクトを指さして"Putthat"と言い、次に目標の場所を指さして"there"と言えば、オブジェクトがそこに移動するというものである。また、調整子の実現例としては、機械による「うなずき」の表示法に関する研究などがある［黒川 88］。

表　情：　表情は情動を表出し伝達する、最も重要なノンバーバルチャネルである。表情が表す基本的な情動は、幸福、悲しみ、恐怖、嫌悪、怒り、驚きの6 つから成る。一方、コミュニケーションにおける表情の重要性が認識されており、電子会議などでの応用を目指して、表情認識・合成の研究が行なわれている。コミュニケーションにおける表情は、身振りにおける調整子とも関連があり、前述のSony CSLの対話的表情によるインタフェースのような研究がある。

視　線：　コミュニケーションにおいて視線は非常に強い力を持つ。明らかにされている視線の機能として、(1)会話の開始合図、主導権の委譲と要求、(2)意思表示、(3)感情表現、(4)関心表明などがある。視線は、アイカメラや顔画像の処理によって検出される。ノンバーバルインタフェースとしての視線の利用は、もっぱら関心の対象の指示や選択に限られているが、世界情報のうち何を中心的に処理すればよいかなどの、いわゆる注視点（focus of attention）の決定に役立つに違いない。

音声などの言語情報は、伝達内容がより厳密ではあるが、そもそも細かいニュアンスや感情の変化などの微妙な情報の伝達が困難である。ただし、音声における韻律情報のうち、どのような成分が怒り、喜び、悲しみなどの感情表現に寄与しているかを分析し、合成音声の韻律を制御することにより、これらの感情を表現する試みが行なわれている［Kitahara 92］。とにかくノンバーバルチャネルを統合したハイパーモーダルインタフェースが実現されれば、人間とコンピュータのつきあい方が大きく変わることは間違いがない。

参考文献

1. [Bobrow 77] D. Bobrow、 R. Kaplan, M. Kay, D. Norman, H. Thompson, and T. Winograd. GUS: A frame-driven dialog system. Artificial Intelligence, 8:155-173, 1977.
2. [Bolt 80] R. A. Bolt. Put-That-There: voice and gesture at the graphics interface. Computer Graphics, 14(3):262-270, 1980.
3. [Chovil 89]N. Chovil. Communicative Functions of Facial Displays in Conversation. Ph.D. Thesis, University of Victoria, 1989.
4. [Ekman 69] P. Ekman and W. V. Friesen. The repertoire of nonverbal behavior: categories, origins, usages, and coding. Semiotica, 1:49-98, 1969.
5. [Ekman 78] P. Ekman and W. V. Friesen. Facial Action Coding System. Consulting Psychologists Press, Palo Alto, California, 1978.

6. [Frijda 86] N. H. Frijda. The emotions. Cambridge University Press, 1986.

7. [Itou 92] K. Itou, S. Hayamizu, and H. Tanaka. Continuous speech recognition by context-dependent phonetic HMM and an efficient algorithm for finding N-best sentence hypotheses. In Proseedings of ICASSP'92, pages I.21-I.24, 1992.

8. [Kernighan 84] B. W. Kernighan and R. Pike. The UNIX Programing Environment. Prentice-Hall, Inc., 1984.

9. [Kitahara 92] Y. Kitahara and Y. Tohkura. Prosodic Control to Express Emotions for Man-Machine Speech Interaction. IEICE Trans. Fundamentals, E75-A(2):155-163, 1992.

10. [Nagao 92] K. Nagao. A preferential constraint satisfaction technique for natural language analysis. In Proceedings of the 10th European Conference on Artificial Intelligence (ECAI-92), pages 523-527, 1992.

11. [Nagao 92(1)] K. Nagao and E. Osawa. A Logic-Based Approach to Plan Recognition and Belief Revision. Technical Report, SCSL-TR-92-007, Sony Computer Science Laboratory Inc., 1992.

12. [Reiter 92] E. Reiter, C. Mellish, and J. Levine. Automatic generation of on-line documentation in the IDAS project. In 3rd Conference of Applied Natural Language Processing, pages 64-71, 1992.

13. [Sherer 80] K. R. Sherer. The functions of nonverbal signs in conversation. in St. Clair, R. N. and Giles, H. (Eds.), The Social and Psychological Contexts of Language, Lawrence Erlbaum, Hillsdale, NJ, pages 225-244, 1980.

14. [Shneiderman 83] B. Shneiderman. Direct manipulation: A step beyound programming languages. IEEE Computer, 16:57-69, 1983.

15. [Stock 92] O. Stock. A third modality of natural language? In Proceedings of the 10th European Conference on Artificial Intelligence, pages 853-862, 1992.

16. [Swagerman 87] J. Swagerman. The Aritificail Concern REalization System ACRES: A computer model that has emotions and can name them. Ph.D. Thesis, University of Amsterdam, 1987.

17. [Takeuchi 92] A. Takeuchi and S. Franks. A Rapid Face Construction Lab. Technical Report, SCSL-TR-92-010, Sony Computer Science Laboratory Inc., 1992.

18. [Takeuchi 93] A. Takeuchi and K. Nagao. Communicative facial displays as a new conversational modality. In Proceedings of ACM/IFIP INTERCHI'93, to appear, 1993.

19. [Weizenbaum 66] J. Weizenbaum. ELIZA-A computer program for the study of natural language commu- nication between man and machine. Communications of the ACM, 9(1):36-45, 1966.

20. [Wilensky 84] R. Wilensky, Y. Arens, and D. Chin. Talking to Unix in English: An overview of UC. Communications of the ACM, 27(6):574-593, 1984.

21. [Winograd 72] T. Winograd. Understanding Natural Language. Academic Press, 1972. Ph.D. Thesis.

22. [Wundt 85] W. Wundt. 身振り語の心理. 福村出版, 1985.

23. [Zeltzer 92] D. Zeltzer. Autonomy, Interaction, and Presence. Presence, 1(1):127-132, 1992.

24. [戸田 87] 戸田正直. 心をもった機械. ダイヤモンド社, 1987.

25. [戸田87(1)] 戸田正直, editor. 認知理論に基づく社会的相互作用過程の解明. 昭和61年度科学研究費補 助金研究成果報告書, 1987.

26. [佐藤 92] 佐藤誠. フォースディスプレイ. テレビジョン学会誌, 46(6):681-684, 1992.

27. [佐藤 91] 佐藤誠, 平田幸広, and 河原田弘. 空間インタフェース装置spidar の提案. 電子情報通信学 会論文誌D-II, J74-D-II(7):887-894, 1991.

28. [小木 92] 小木哲朗and 廣瀬通孝. 風覚ディスプレイによるデータの表示. In 第8 回ヒューマン・イ ンタフェース・シンポジウム, pages 413-418. 計測自動制御学会, 1992.

29. [長尾 92] 長尾確and 丸山宏. 自然言語処理における曖昧さとその解消. 情報処理, 33(7):746-756, 1992.

30. [沼崎 91] 沼崎俊一, 佐藤誠, 平田幸広, and 河原田弘. 空間インタフェース装置spidar を用いた回転 形状モデリング. 信学技報PRU91-71, 91(298), 1991.

31. [楠見 90] 楠見孝. ヒューマン・インタフェースにおけるメタファ・アナロジーの役割. 日本認知科学 会R&I 研究会, SIGR&I90(2), 1990.

32. [黒川 88] 黒川隆夫. ヒューマン・インタフェースとしての動作言語. 計測と制御, 27:49-55, 1988.

33. [服部 91] 服部桂. 人工現実感の世界. 工業調査会, 1991.

2.3 話者認識

2.3.1 はじめに

音声による個人の認識は、バンキングやデータベースへのアクセスにおける鍵としての利用をはじめ、音声認識装置の特定話者の選択的応答等、音声認識システムと組み合わせた多くの応用の可能性がある。

話者認識に関する研究の歴史は古く、最初は犯罪捜査への利用を目的として研究がはじめられた。初期の研究は、聴覚やソナグラフに基づく視覚によるものであったが、その後、計算機が利用できるようになって、自動認識の研究がはじめられた。初期の研究では、話者を特徴づける音響パラメータや、発話内容を限定した認識手法が検討されたが、最近ではベクトル量子化やHMM等の新しい手法を用いた話者認識、特に発話内容を限定しない認識手法に大きな進展が見られる。これまでの研究に関しては多くのレビューがあるので[Hecker 71][Atal 76][新美 79][古井 85][Doddington 85][O'Shaughnessy 86][古井 91][Naik 90][Furui 90]、以下ではこれらを参考に、話者の自動認識に関する現状と今後の課題について述べる。

2.3.2 話者認識系の構成

（1）話者認識の形態

話者認識には、未知話者があらかじめ登録された話者のうちの誰の声であるかを判断する話者識別（Speaker Identification）と、未知話者が本人であるかどうかを判定する話者照合（Speaker Verification or Authentication）の2つの形態があり、これらを区別しない場合は単に話者認識と呼んでいる。話者照合は、指紋や印鑑と同様にバンキングや、データベースのアクセス、各種機器のアクセスの制御のように、指紋やkeyとしての応用が考えられ、話者識別は、音声タイプライタによる議事録作成のような応用で、話者の自動ラベル付のようなタスクに応用できる可能性がある。

話者識別ではあらかじめN人の話者の標準パターンを登録しておき、未知話者の音声が入力されると、特徴ベクトル系列を抽出し、標準・テストとの距離（又は類似度）を計算し、最も近い標準パターンの話者を入力の話者と判定する（図2.3-1(a)）。したがって、識別の精度は、単語認識と同様、正答率や誤り率で評価される。また、目的によっては、登録話者外のカテゴリーを儲けリジェクト機能を持たせる場合もある。一方、話者照合では、照合すべき入力話者の音声パターンと、あらかじめ登録されていた本人の特徴パターンとの距離を計算し、これがある閾値より小であれば本人と判定し、大であれば詐称（Imposter）の声であると判定する（図2.3-1(b)）。したがって、話者照合では次の2種類の判定誤りがある。

① 本人否却率（False Reject Rate）：本人の声を詐称者の声と誤る割合

② 詐称者受理（(False Acceptance；FA)：詐称者の声を真の話者と誤る割合

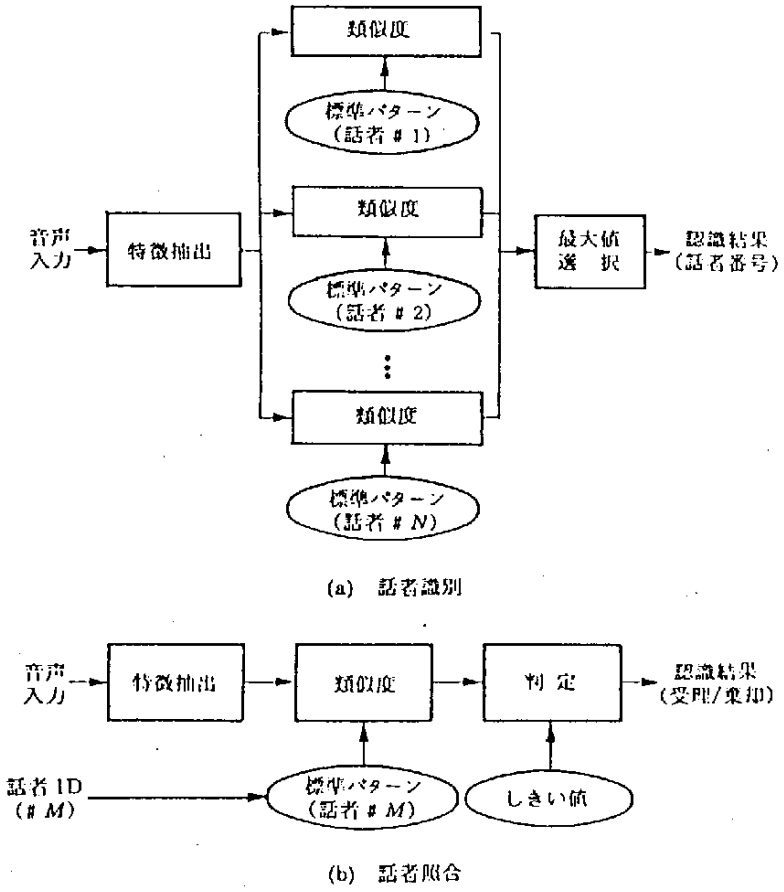


図２．３-１　話者識別と話者照合の構成［古井 91］

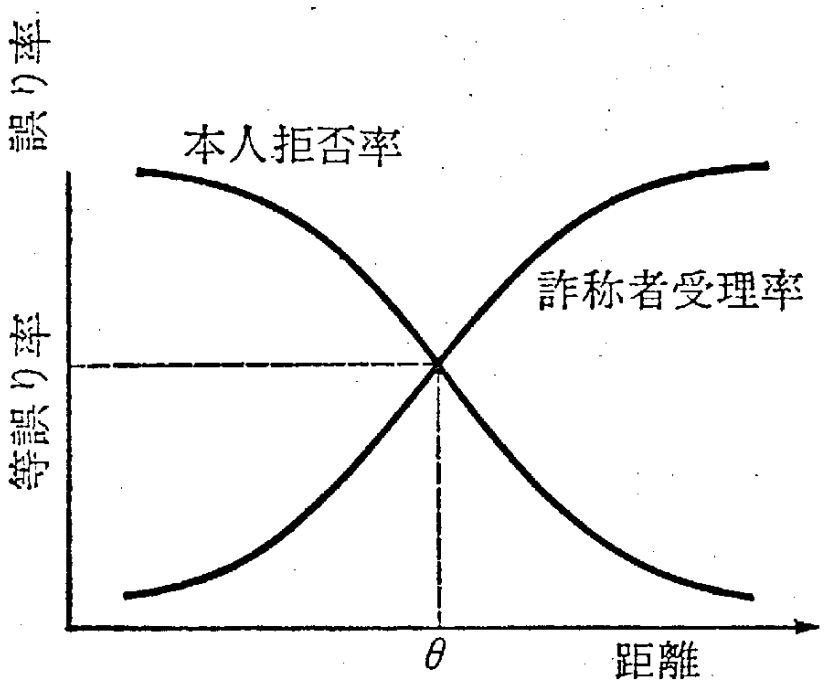


図２．３-２　話者照合における誤り率の定義

これらは、図2.3-2に示すように判定の閾値を小さくすれば、詐称者受理率は減少するが、本人否却率が増大するという関係がある。そこで、事後的に両誤りが等しくなるように閾値を設定したときの誤り率を等誤り率（Equal-error-rate）と呼び、通常はこれで話者照合システムの性能を評価する。実際上は、どちらの誤りを重要視するかは応用により異なり、たとえば、詐称者受理誤り率をある割合に設定したときの本人否却率で評価する場合もある。

一般に、話者識別と話者照合の認識性能は、条件が同であれば、後者の方が良い。認識対象とする話者数と話者識別および話者照合の誤り率の関係を比較した例を図2.3-3に示す[Doddington 85]。この結果は、各話者の特徴ベクトルが、対角等分散の多次元ガウス分布をし、また、全話者の平均ベクトルが平均0で各話者の分散のS倍の対角等分散を示すという仮定のもとで、シミュレーションにより得られたものである。これからわかるように、話者識別では、話者数が増加するにしたがい誤り率は1に近づくが、話者照合では、話者数の認識精度への影響は小さい。実際の音声データを用いた認識実験でも同様の結果が得られている[古井77]。

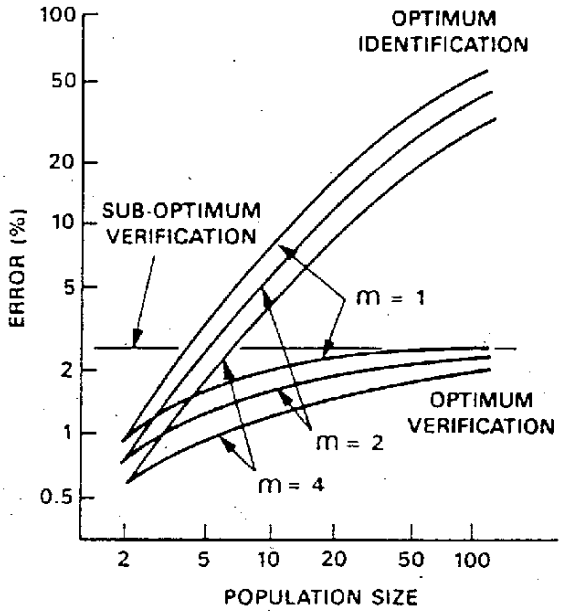


図2.3-3　話者識別と照合における誤り率と話者数の関係［Doddington 85］

（2）テキスト依存・独立型話者認識

音声には、言語情報、話者情報、情緒性情報等が複雑に絡み合っているため、話者情報のみを分離抽出し話者認識する方法はまだ確立されていない。そこで、言語情報と話者情報を分離するかわりに、認識に用いる音声の言語内容を固定し、同一の単語や句の話者間の差を計算し

て認識をおこなうのが最も簡単で、これをテキスト依存型話者認識と呼ぶ。しかし、人間の聴覚では、知っている人の声であれば、発話内容にかかわらず話者の認識が可能である。これと同様の認識を計算機で実現しようとするのがテキスト独立型話者認識である。この場合には、話者間の類似度の導出で、発話内容の影響をいかに抑えるかが大きな課題となる。一般に、テキスト依存型の方が独立型に比べ認識精度が高く、初期の研究は、主にテキスト依存型であったが、最近、テキスト独立型の認識手法に大きな進展が見られる。

バンキングやデータエントリのKey として話者認識を応用する場合には、テキスト依存型で目的は達成されるが、複数話者のもとで特定の話者のみに選択的に応答するような音声認識装置を実現したり、複数話者の音声が含まれるデータベースから特定の話者の音声のみを検索するような応用では、テキスト独立型の認識が必要とされる。また、ユーザにとって使い易いシステムとするためには、やはりテキスト独立型の方が好ましい。

2.3.3 話者の音響的特徴

音声の個人性の音声生成過程における要因は、①発声器官の解剖学的な個人差、②音声の学習過程の個人差の2つに分類することができる。①は主に声道の大きさ、形状、声帯の長さ、厚さなどで、音響的にはホルマント周波数や、ピッチの平均的な違いとして表われる。一方、②は方言や話し方の個人差など、ホルマント周波数、ピッチ、振幅（パワー）の時間的変化パターンなどにその違いが現われる。以上のように、解剖学的差異は主に静的な特徴量に、また、話し方などの個人差は動的な特徴量に反映される。これまで、話者の特徴を担う音響パラメータとして、多くのものが検討されてきたが、以下にその概要を述べる。

（1）話者認識に有効な音声区分と音響パラメータ

音声の個人差は、音韻や音声区分により一様ではないため、話者認識に有効な区分とその音響パラメータを見出すことが重要である。話者認識に有効な音響的パラメータを見出すために、これまで、

1）F比（分散比）による方法
2）Divergenceによる方法
3）認識誤りによる方法

の方法が提案され、膨大なデータを用いた検討がいろいろなされている。

1）は、単一パラメータに対して、話者間の分散と話者内の変動による分散の比が大きいパラメータを選択するもので、最も単純な方法でよく用いられてきた。また、複数パラメータの

線形結合でパラメータを導く場合も、F比を最大化することにより固有方程式が導かれ、最適な結合係数を導くことができる。2）の方法は、確率分布間の距離としてKullbackによって導入されたもので、1次元の特徴パラメータに対するF比を、多次元ベクトルに拡張したものに相当する。

初期の研究でWolfは、主にフィルタ群分析で得られる種々の音声セグメントの特徴をF比をもとに比較評価し、重要なパラメータとして、①基本周波数、②母音および鼻音スペクトルの特徴、③音源スペクトルの傾斜、④単語の継続時間、⑤ボイスオンセット時間、等を見出している。上位17のパラメータを用いて認識実験をおこない、21名の男性話者に対し、話者識別で誤り率0％、また話者照合で誤り率2％を得ている[Wolf 71]。

以上の1）と2）で得られるパラメータは、実際の認識系で用いたときの認識誤りとは直接的な関係はない。しかし、特徴パラメータに正規性の仮定をし、Bayesの決定法を適用すると誤り率の上限を計算できる。そこで、N個のパラメータのうち、任意の1パラメータを除いたときのこの誤り率を計算し、最も誤り率の小さいN－1個のパラメータを求めるという方法が提案されている。これをKnock-out法と呼び、Samburは種々の音声セグメント中のいろいろの音響パラメータの話者識別への有効性の比較をおこなった[Sambur 75]。

表2.3-1　各種音響パラメータの話者認識における貢献度の順位［Sambur 75］

Ordering of Features

Feature	Speech Event	Feature	Speech Event
1. NF2	/n/	20. THISFO	FO
2. UF3	/u/	21. MANF2	/m/ in man
3. IF2	/I/	22. MANB3	/m/ in man
4. K	duration of /k/	23. EEF1	/i/
5. REMF3	/m/ in remember	24. EEF4	/i/
6. NF6	/n/	25. EEF3	/i/
7. REMF4	/m/ in remember	26. SHF2	/sh/
8. CASHFO	FO	27. AEF2	/ae/
9. IF4	/I/	28. AEF4	/ae/
10. AI	F2 slope in /aI/	29. AEF1	/ae/
11. REMFI	/m/ in remember	30. SF2	/s/
12. AVFO	FO	31. UF4	/u/
13. SF3	/s/	32. IF1	/I/
14. UF2	/u/	33. BONDFO	FO
15. EEF2	/i/	34. REMF6	/m/ in remember
16. NF1	/n/	35. IF5	/I/
17. MANF4	/m/ in man	36. MANB4	/m/ in man
18. UF1	/u/	37. AEF3	/ae/
19. NF3	/n/	38. SHF1	/sh/

実験では、22～42才の11名の話者の３年にわたる音声データを用いて、ＬＰＣ分析で得られる種々の音響パラメータの有効性を比較している。これらの結果（表２.３-１）から、話者認識に最も重要な音響パラメータとして、①鼻音/n/の1000kHz付近の第２共振周波数、②鼻音/m/の第３、第４ホルマント周波数、③母音の第２、第３、第４ホルマント周波数、④平均基本周波数、等を見出している。また、最良の５つのパラメータを用いて話者識別をおこなった結果、320回の識別で誤りは１回のみであった。

これらの特徴パラメータの抽出は、人手によるセグメンテーションを必要とするので、自動認識には不適当であるが、音声のどのような言語内容のどの音響パラメータに有効な個人差が含まれているかを示す基礎データとして重要である。

以上の例に見られるように、母音や鼻音のホルマントは優れた認識率を与える。また、母音、鼻音、摩擦音スペクトルは

1) 音声信号内で容易に同定できる
2) 話者の識別に有効な特徴を含んでいる

という点で有用である。特に、鼻音は鼻腔の個人差が大きく、風邪のときを除けば意識的にかえることが難かしいので話者認識に適している[Glenn 68]。

（２）話者認識に有用なＬＰＣパラメータの比較

線形予測分析で得られるパラメータとして表２.３-２に示すものがある。

表２.３-２　各種LPCパラメータによる話者識別精度

LPCパラメータ	認識率
ケプストラム係数	70.3　%
線形予測係数	63.8
自己相関係数	59.7
対数面積関数	57.0
インパルス応答	60.6
反射係数	――

これらパラメータの話者認識における有効性は、類似度（相違度）にどのような尺度を用いるかにも依存するが、代表的な比較評価実験例として、Atalらの研究がある。この研究では、10名の女性の発声した"May we all hear a yellow lion roar ?"という文を、標本化周波数10kHzで12次のＬＰＣ分析を行い、各発話の無音およびポーズ除いた区間を等分割した時点の40フレームのパラメータを抽出している。全話者、全フレームの平均話者内分散を共分散とするマハラスビス距離を用い、各フレームに対して話者識別をおこなった結果、表２.３-２に示すように、ケプストラム係数が70.3% で最も良く、面積関数が57·0% で最も認識精度が低いことがわかった。また、複数フレームに対する平均距離で認識した場合、ケプストラム係数を用いると、0.5s相当のセグメントの認識率は98% であった[Atal 74]。この他、Shridher らは、表２.３-２のＬＰＣパラメータの他に、対数ＬＰＣスペクトルおよび16 BPFスペクトルを取り上げ、これらをＫＬ展開した特徴空間における重み付きユークリット距離を用いて、テキスト独立型話者認識をおこなった[Shridher 82]。この実験でも、情報量の多いＬＰＣスペクトル（33係数）を除くと、ケプストラム係数が話者識別と照合の両方に対して良い結果を与えている。

最近では、ＬＰＣケプストラム係数が、対数スペクトルを表すパラメータとして最もよく用いられる。また、調音結合等のミクロな動的個人差を表現するパラメータとしてこれらの時間変化パターンの一次または２次回帰係数がよく用いられる。

（３）プロソディーに関する特徴

通常、音声認識や話者認識では主にスペクトルの特徴が用いられるが、話者認識ではプロソディー（特に基本周波数）も重要である。テスト音声の平均基本周波数は、成人男性、成人女性、子供のような話者の大まかな分類に有効である。一方、F_0の動的特徴（時間変化パターン）は時期による変動は大きいが、テキスト依存型の認識では有効な特徴である。たとえば、Atal は、各発話の基本周波数パターンを1／40の等間隔で標本化し、K-L 変換で20に圧縮（全分散の99.5%）した特徴量とマハラノビス距離を用い、10人の話者識別で97% の認識率を得ている[Atal 72]。しかし、同じ音声を12次のＬＰＣケプストラムで認識した場合、0.5s長の音声で98% の認識率が得られるので、スペクトルの方がずっと有効であるといえる。また、Sambur(1975) の実験では、文中の特定部分のF_0パターンの傾斜を動的特徴として用いたところ、時期的変動（３年半にわたる５セッション）が大きく、信頼できる特徴にはなりえなかったことが明らかにされている。

F_0はまた、パワーパターンと組合せて使用される場合がある[Rosenberg 76]。一方、時間的特徴、例えば、音韻の継続時間、発話速度は話者の特徴になりうることは容易に想像できるが、これらの特徴はあまり使われたことがない。Johnson の研究では、単に"Speech burst"（エ

ネルギーがある閾値を越える区間）に関する40の統計的タイミング測度を用いることで良い結果を得ている[Johnson 84]。テキストが限定されている場合、発声スタイルはそのようなバーストパターンに反映する。他の研究では、単語長、ボイスオンセットタイムの有効性が示されている[Wolf 72]。

（4）時期的変動

同じ話者が同じ内容の音声を発声してもその音声は変動する。特徴量の継時変化は音韻による変動に比べ十分小さいので、音声認識ではあまり問題とならないが、個人性の情報は音韻性情報に比べ細かいため、話者認識では問題となる。このことは最初 Luck により指摘され[Luck 69]、古井らによって詳細な検討がなされた[古井 74]。その結果、音声の継時変化の主な要因は、音源の時期的変動であることが解明されている。

時期的変動の影響を低減するするためには、学習サンプルに長期間にわたるデータを用いたり、標準パターンを適当な周期で更新するのが有効であるが、この他に、音声の時間平均スペクトルを１次系または２次臨界制動系で近似し、この逆フィルターでスペクトル等化処理をおこなう方法が提案されている[古井 74]。たとえば、単語音声スペクトルの統計的特徴を用いた話者照合において、各話者が短期間（10日間程度）の間の４日間に発声した12サンプルを学習サンプルとし、６週間から10カ月後のテストサンプルを照合した場合、スペクトル等化により誤り率を１／３程度に低減できることが明らかにされている。

2.3.4　テキスト依存型話者認識

（1）ＤＴＷによる方法

テキスト独立型話者認識では、発話の時間軸の変動を吸収するのに、単語認識と同様、動的計画法による時間軸の伸縮（ＤＴＷ）がよく用いられる。一例として、古井らの方法では、10ｍｓのフレーム周期で１～10次のＬＰＣケプストラム係数を抽出したあと、全発話（短文）の平均ケプストラム係数を各フレームのＬＰＣケプストラム係数から差し引き、伝送系の周波数特性や話者内のグローバルなスペクトル変動を除去し正規化する。さらに、動的特徴として、ケプストラム係数の時間関数に多項式を当てはめ、１次および２次のデルタケプストラムを10ｍｓ周期で抽出し、正規化ケプストラム係数と合わせ20次元の特徴ベクトル系列とし、標準・テストパターン間の距離をＤＴＷにより計算し、閾値と比較し話者照合の判定をする。実際に、６か月にわたり実験をおこない、男性60名、女性60名に対し、平均認識率97％を得ている[Furui 81]。

（2）ＨＭＭによる方法

ＨＭＭはスペクトルの統計的変動を効率良くモデル化できるため、ＤＴＷによる方法より高い認識精度が期待できる。話者認識にＨＭＭを利用した方法がいくつか試みられている[Zheng 88][Naik 89]。また、Rosenberg らは、発話をサブワードのＨＭＭの連結であらわし、話者を照合するシステムを提案している[Rosenberg 90]。サブワードとしては、擬音韻（ＰＬＵ）と音響セグメント（ＡＳＵ）の２種類の単位について検討している。ＰＬＵは発話の音声表記に基づいて組み合わされ、一方ＡＳＵはなんら言語学的知識を用いずに音響信号から直接導かれる。100人の話者の20000の孤立数字音声を用いて照合性能を評価した結果、ＰＬＵとＡＳＵの性能にはわずかの差しかないことが分かった。結局、照合誤り（等誤り率）は１数字音声で７～８％、７数字音声を用いた場合で１％かそれ以下であった。

このように、ＨＭＭはサブワードモデルの連結により任意の単語や文を合成して、認識に利用できるという利点はあるが、学習に大量のデータを必要とするという欠点がある。

２.３.５　テキスト独立型話者認識

音声に含まれる情報を、言語情報と個人性情報に分離抽出することは難しい。音響パラメータの中には、基本周波数のように言語情報と比較的独立性の高いものもあるが、スペクトルパラメータは音韻による変動が非常に大きいため、テキスト独立型話者認識では、発話内容の影響を如何に低減するかが重要な課題である。

（１）長時間平均による方法

発話内容の影響を低減する最も簡単な方法は、特徴量の長時間平均を取ることである。標準・テスト音声中に含まれる音韻の生起確率がほぼ同じになる程度の時間長で平均をとれば、標準・テストパターン間の音韻系列の差の影響は小さくなると期待される。古井らは、10～30ｓ長の長時間平均スペクトルのケプストラムを用いて９名の話者認識をおこない、10ｓの平均で話者識別で88～89％、話者照合で94～95％の認識率を得ている[古井 72]。

通常話者認識では、読みあげ文章を音声データに用いることが多いが、Markelらは、インタビューで録音した発話内容の自由な音声データを用いて、長時間平均特徴量による話者認識をおこなっている[Markel 77]。３か月にわたって録音した男性11名、女性６名の音声に対し10次のＬＰＣ分析をおこない、ピッチ、反射係数およびそれらの分散を特徴量として、判別分析（話者間・話者内分散比最大）をおこない、その線形直交変換ベクトルを平均化して認識をおこなった。40ｓの音声で話者識別率98％、照合率96％の認識精度を得ている。

この他、Shridharらは、各種のＬＰＣパラメータをＫＬ展開した特徴ベクトルを用い、２種類の特徴ベクトルによる認識結果を論理的に組合せることで、12名の男性に対して、５ｓの音声から99％の話者識別率を得ている[Shridhar 82]。

（2）音韻区分の検出による方法

任意のテキストから話者を認識する直接的方法として、入力音声中の特定の音声区分を検出し、その標準パターンと比較して認識する方法が検討されている[Pfeife 79]。しかし、この方法では、特定の音韻や音声区分を検出または認識するときの精度が話者認識精度に大きな影響を与えるという欠点がある。また、Shwartz らは、音韻検出の問題を避け、短時間の言語情報から認識する方法を提案した[Shwartz 82]。この方法では、短時間の特徴ベクトルは標準話者に関して統計的に特徴づけ、この統計的モデルに基づいて入力音声の尤度を計算し認識する。この方法は歪みや雑音を含む音声に対しても適用され、ある程度よい結果が得られている[Krasner 84]。

（3）音韻のクラスタリングによる方法

松本らは、発話中の特定音韻を検出するときの誤りの影響を避けると同時に、入力音声中の全ての個人性情報を無駄なく利用し、短時間の任意テキストの音声から話者を認識するため、部分空間毎に正準判別分析を適用する方法を検討した[Matsumoto 89]。この方法は、音声の個人性情報が広範囲の音韻に共通するものと、各音韻に依存した情報から成り立っているという考えに基づいている。そこで、図2.3-4に示すように、まず5母音と2音節中の3鼻音から導かれた音韻判別空間で8つの音韻に対応する部分空間に学習用スペクトル系列（20次ケプストラムおよび基本周波数）を分類する。次に各部分空間ごとに正準判別分析を適用し、話者間対話者内分散比を最大化する話者空間を導き各話者の分布をガウス分布で近似する。認識時には、未知音声の各特徴ベクトルを最小距離に基づいて8音韻に対応する部分空間に分類し、その話者空間における各標準話者に対する尤度を計算する。

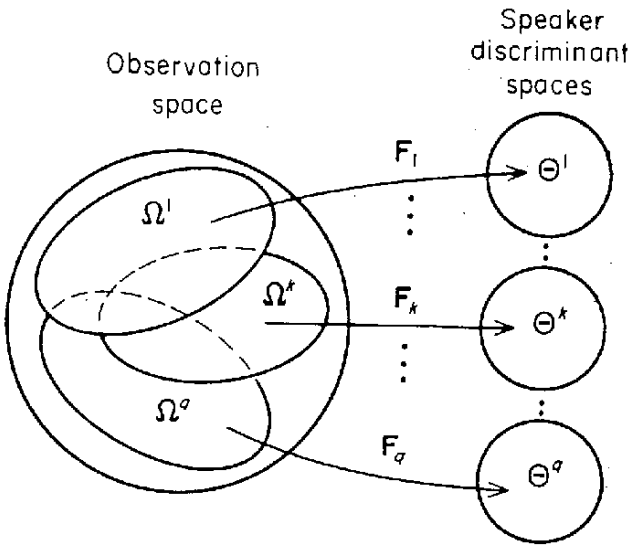


図2.3-4　断片的判別分析に基づくテキスト独立型話者識別の構成

これを全ての入力特徴ベクトルについておこない全尤度（同時確率）を計算し最大尤度を与える標準話者を入力話者として判定する。この方法では各部分空間は音韻を正しく認識することが目的ではなく、音韻による変動を抑えかつ共通の個人性情報が含まれるように分類することが目的である。各部分空間の大きさ（半径）を調べた結果、前舌母音や鼻音ではやや小さめが良いが、一般に複数の音韻を含む程度のかなり重なり合った大きさが適当であることが明らかにされている。

学習音声とテスト音声が同時期（２カ月間の３時期）である場合、40ｍｓ、0.6ｓ、1.4ｓの有声音声による男性10名の話者識別率は65％、98％、100％であった。また、学習音声とテスト音声が半年離れている場合、0.5ｓ以下の有声音声では同時期に比べ認識率は約10％低下するが、1.4ｓあれば100％の認識が可能であった。また、この方法は空間を分割しない場合に比べ学習データは少なくてすみ、１話者・１時期あたり４ｓで認識率は飽和することが明らかにされている。この実験では、部分空間の設定に話者の学習とは別の母音と鼻音データを用いているが、話者空間の学習に用いるデータに対してベクトル量子化や次に述べるエルゴードHMMを用いて分類すれば、より最適な部分空間が導けるものと期待される。

また、Liらは各話者ごとに100ｓの会話音声をＬＰＣ分析し、得られたスペクトルパターンの集合をクラスタリングし、40種類の代表パターンを作成した。認識では、未知話者の音声を短時間ごとにＬＰＣ分析し、得られた全フレームのスペクトルパターンを各標準者の40パターンに対する最小距離を計算し、この距離分布が最も小さい方に偏っている標準話者を入力話者の音声であると判定する。10、５、３ｓのテキストを限定しない未知音声に対して、男性11名の話者識別で、それぞれ96、87、79％の認識率を得ている[Li 83]。

（４）エルゴードＨＭＭによる方法

テキスト独立な話者認識で、音韻に依存した個人性の特徴を利用するため、エルゴードHMMを用いて自動的に大まかな音韻の分類をおこなう方法が提案されている[Poritz 82]。Savicらは、各話者の学習音声を５状態のエルゴードＨＭＭでモデル化し（図２.３-５）、未知音声を各話者のＨＭＭを用いて各状態に区分化している[Savic 90]。

５つの状態は、大まかに２つの母音カテゴリー、鼻音、摩擦音、破裂音に対応している。次に、各状態に区分化されたフレームから特徴を抽出し、学習時に抽出されたテンプレートとマハラノビス距離を計算し、閾値と比較してスコアを求める。最終的な照合は、各カテゴリーのスコアの線形結合によって判定する。実験の結果、各カテゴリーにより話者照合への有効性が異なるため、このスコアの線形結合により認識精度を改善できることが示されている。この他、エルゴードＨＭＭの尤度を用いた認識も検討され、次に述べるＶＱによる方法と同様、認識精度はモデルに含まれるガウス分布の数に依存することが確かめられている[松井 91]。

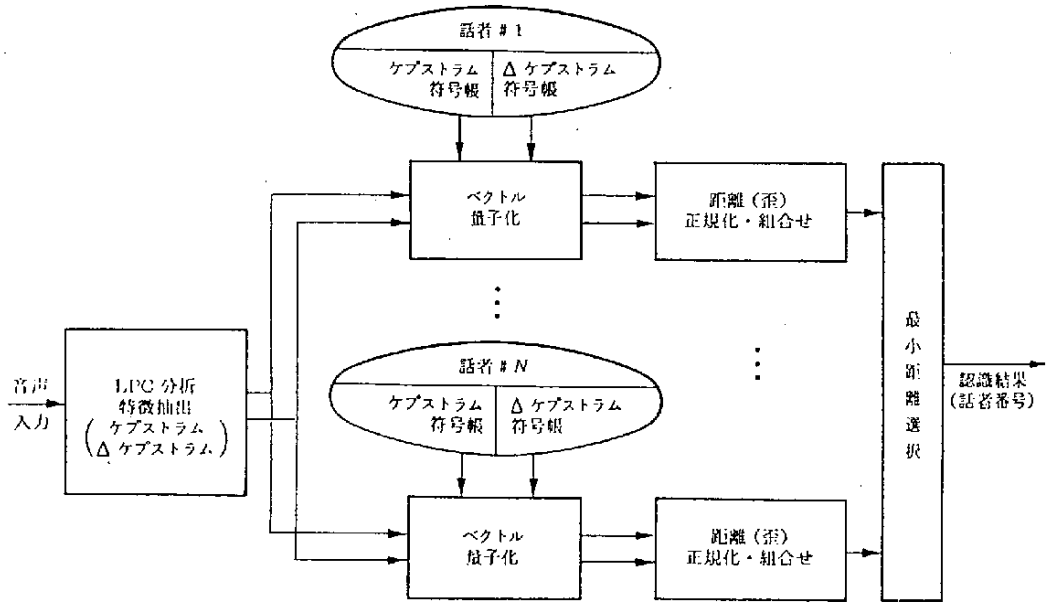


図２.３-５　ベクトル量子化に基づく話者認識の構成［古井91］

（５）ベクトル量子化による方法

Soong らは、ケプストラムとデルタケプストラムの２つのコードブックによる同様のテキスト独立型認識実験をおこなった[Soong 85]。音声データとして孤立発声の10数字音声を用い、各話者ごとに10数字を用いてコードブックを作成する。認識時には、任意の数字音声をテスト音声として２種類のコードブックに対する平均ＶＱ歪みを求め、両歪みを最適に組合せて認識する（図２.３-６）。

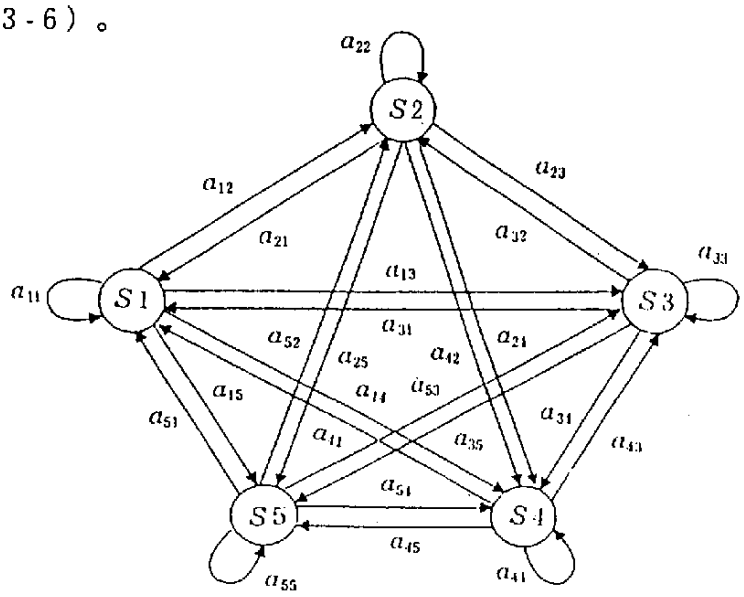


図２.３-６　エルゴーディックHMMの例［Savic 90］

実験の結果、ケプストラムとデルタケプルトラムは比較的相関が小さく、認識精度の向上に寄与することが確かめられ、また、デルタケプストラムは伝送歪みの影響を受けにくいことが明らかにされている。なお、この実験は、語彙が数字音声に限定されているので、完全なテキスト独立型ではない。

また松井らは、テストデータの中の符号帳の分布から外れた部分の影響を避けるため、通常のＶＱ歪みの代わりに、テストデータの分布と符号帳の分布の重なりを考慮した「歪み・重なり尺度（ＤＩＭ）」と呼ぶ歪み尺度を提案している[Matsui 90]。特徴ベクトルの成分として、１～16次のケプストラムとピッチ（有声部の場合）、および両方の線形回帰係数を統合して１特徴ベクトルとして表し、ＤＩＭによる話者認識をおこなった。この方法では、パラメータを話者内変動を抑え話者間変動を強調するように変換した後、有声部と無声部に対して、それぞれ別々の符号長を作成し、認識時にはそれぞれ対応する符号長で量子化したときのＤＩＭを計算し、両者の歪みを加えて判定をおこなっている。男性９名が３年間にわったて発声した音声を用い、一時期の単語と文章の音声で符号帳を作成し、異なる時期に発声した30秒の文章音声で認識をおこなった。その結果、話者識別で99.0%、話者照合で98.7%の平均認識率を得、特徴量の正規化法とＤＩＭの有効性を示している。

この他に、ベクトル量子化歪みの値ではなく、歪みの分布の違いにより話者を認識する方法も提案されている[白井 87]。

（６）確率分布による認識

この方法は正統的なパターン認識手法で、学習データにより各話者の特徴ベクトルの確率分布を求めておき、認識時には未知音声の各標準話者に対する尤度を計算し、話者を決定する方法である。Sebestyen は、第１～第４ホルマント周波数を含む13のパラメータで20ｍｓごとの音声の特徴を表し、この確率分布をクラスタリングの手法で等分散ガウス分布の混合で近似した。男女を含む11名の音韻バランス文を用いて種々の実験をおこなっているが、話者照合の実験では、１話者の２文で確率分布を学習し、１ｓ長（５フレーム）の音声に対する棄却率が10%になるように閾値を設定した場合、詐称者受理誤りは７％であった。また、学習外の１ｓ長の音声に対する本人棄却率は３％であった[Sebestyen 62]。この実験は、分析にソナグラフを使うなど分析データの信頼性に問題はあるが、テキスト独立型の話者認識に関する最初の研究であるだけでなく、ＶＱによる話者認識がこの手法の近似になっているという点で重要である。

最近、RoseらはＥＭアルゴリズム（The Expectation Maximization）で混合ガウス分布を導き、12名の話者の識別実験で、１ｓの会話音声から89%の認識率を得ている[Rose 90]。Tseng らも、ＥＭアルゴリズムで学習した混合ガウス分布（この分布をthe Continuous

Probabilistic Acoustic Map:CPAMと呼んでいる）によるテキスト独立な話者認識をおこない、ＶＱによる方法との比較を行っている[Tseng 92]。８次のＬＰＣケプストラムとデルタケプストラムを組合せた16次元特徴ベクトルを用い、10数字音声による20名の話者（男女各10名）の認識をおこなった結果、ＶＱに基づく方法の誤り率は、照合が 9.70%、識別が 4.87%であるのに対し、ＣＰＡＭではそれぞれ 8.01%, 2.86% で、ＶＱ法より優れていることが確認されている。

（７）ニューラルネットによる方法

テキスト独立型話者認識においても、多層状ニューラルネットワークが利用されている[Rudasi 91]。話者認識では、普通３層ネットワークが用いられ、出力ユニットが各話者に対応付られる。学習時には全話者の学習サンプルを用いて結合係数を決定する。認識時には未知音声の各フレームをネットワークに入力し、各出力ユニットの出力値の累積が最も大きいユニットに対応する話者を入力話者と判定する。この他に、話者認識で用いられるニューラルネットワークとして、予測型のネットワークがある。この場合には、過去の数フレームから次のフレームを予測する非線形予測子にニューラルネットワークが用いられる[Hattori 92]。学習時には各話者に対して予測誤差が最小となるように学習する。認識時には各話者の予測ニューラルネットワークを用いて予測誤差を計算し、最小予測誤差を与えるネットワークに対応する話者を入力話者と判定する。このとき、全ての音声フレームに対して同じ予測子を用いると予測誤差が大きくなるので、クラスタリングによりいくつかのグループ（状態）に分けて各グループごとに異なる予測子を用意する方法が提案されている。服部はＴＩＭＩＴデータベースから24名の女性話者を選び、聴覚モデルで分析された40チャンネルのMean Rate Response を平均０、最大1.0に正規化した値を特徴ベクトルとし、ＶＱ法、ＨＭＭ法、多層状ニューラルネットワーク等との性能を比較した。その結果４状態の非線形予測型ニューラルネットワークが最も認識率がよく、３秒のテスト音声で100%の認識率を得ている。

２.３.６　実際上の問題

（１）多段階認識と逐次判定法

話者識別では話者数に比例して計算量と応答時間が増加する。この計算量の増大を避ける１つの方法として、話者を階層的にクラスター化しておく方法がある。たとえば、基本周波数を使って話者を比較的少数のグループに分類することが可能である。

また、話者照合の実際の応用では、話者の受理／棄却の応答を実時間でおこなう必要がある。そのようなシステムでは、逐次判定法が有用である。この方式では、判定の境界付近のテスト入力に対しては認識を確定せずに次の入力まで判定を保留する。そのため２つの閾値を用いて

次の3通りの判定をおこなう。

1) 低い方の閾値以下なら受理する
2) 高い方の閾値以上なら棄却する
3) 2つの閾値の間にあるときは判断を保留し、再入力を要求する

このようなアプローチをとることにより、類似する話者間の誤りを避け、テスト用の発話を短縮し、応答を高速化できる。

たとえば、Rosenberg は、300 ～ 3000 Hz帯域の一様 17ch BPFスペクトルを使い、話者照合に逐次判定法を適用した。テンプレートは、4つの単音節の各々から最大エネルギー時点の回りの20ms間隔毎の6フレームからなる17次元ベクトルで構成される。50名の話者と70名の詐称者を用いた場合、1つのテストフレーズで誤り率 1.6%、第2、第3のフレーズを入力した場合、誤りは0.42%, 0.23% に減少した。逐次判定により平均 1.3フレーズ、0.3% の棄却誤り、 1% の受理誤りを得ている[Rosenberg 76]。

（3）伝送特性の影響

電話回線、特に市外回線を通った音声は、伝送系の変動が大きいため種々の歪みを受ける。これらのうち①減衰歪み、②回線雑音、③非直線歪み、が音声認識や話者認識で問題となる。減衰歪みは回線により異なり、また300 ～3200Hzの急俊な帯域ろ波特性を受ける。したがって、電話音声に対する話者認識システムの評価には、標準・テスト音声として異なる回線を通したものを用いる必要がある。

減衰歪みに対処する方法として、市川らは、帯域選択自己平均逆フィルタ法と名付けた方法を提案し、ｃｏｓｈ尺度を用いて25名の話者で86％の照合率を得ている。この方法は、回線による減衰歪みの影響を低減するため、標準、テスト単語をれぞれの平均パワースペクトルを計算し、その逆スペクトルを各フレームに乗じて伝送特性の成分を除去している。また、300 ～3200Hz外の帯域の影響を低減するため、この帯域に対して選択的線形予測分析を適用している。また、非線歪みの影響を受けにくいF_0の距離をｃｏｓｈ尺度と組合せることで、照合率は65％から91％に改善されることを示した[市川 79]。

（4）符号化音声による認識

ディジタル符号化音声では、品質低下の影響も問題となる。Everett らは、種々のタイプの4kHz 符号化音声によるテキスト独立型話者認識を試みている[Everett 85]。20名の話者について、10次の LPC 反射係数と、10次のケプストラ係数を用いて評価したところ、符号化し

ないもとの音声の照合率が 95%であったのに対し、2.4kbit/s の LPC符号化音声では、認識率は80%に低下した。波形符号化では量子化雑音のために認識率はかなり低い。ゲインは、電話系や符号化(LPC or ADPCM)の歪みの影響を受けにくいように思われるが、ゲインはむしろ性能を低下させるという結果もある[市川 79]。

（5）雑音環境における話者認識

雑音環境下の話者認識に関する研究はまだあまりなされていない。Openshaw らは、話者識別（20名）における種々のパラメータの性能比較をおこなっている[Openshaw 92]。認識には、コードサイズ 32 の量子化誤差を用いている。 Mel, PLP とその動的パラメータ、RASTA 処理、NormおよびAngle歪み尺度、線形判別分析による特徴量の組み合わせについて、白色雑音(15dB)およびスペクトル傾斜の歪みに対する耐性を調べている。その結果、①動的特徴は高SNR で静的特徴より認識精度が低いが、低 SNR で良い、②PLP はスペクトル傾斜の変動に弱いが、RASTA 処理により大きな改善が得られる、③雑音付加データを学習用音声に含めることで(50%)、耐雑音性が向上する、④パラメータの組み合わせの中で、PLP-RASTA と デルタメルケプストラムの組み合わせの方が、メルケプストラム + デルタメルケプストラムよりも優れていることが明らかにされている。

この他、群遅延から抽出したピッチ情報を使い、雑音に強い単語境界仮設を導き、それに基づいて話者認識をおこなう方法が提案されている[Yegnantiayana 92]。

2.3.7 おわりに

話者認識に関する基本的事項と、これまで提案されてきた話者認識手法について概説した。現在、話者照合の精度は、テキストを限定すれば、暗唱番号等と組み合わせることにより、適用分野によっては実用可能なレベルにあると思われる。しかし、バンキングにおける鍵としての利用のように高信頼性が要求される応用では、一層の精度向上を図るとともに、雑音や伝送系の歪みに頑健な認識手法を開発する必要がある。また、最近、テキスト独立型の話者認識技術が大きく進展しているが、話者ソーティングや音声認識における特定話者の選択的応答等への応用では、さらに短時間音声に対する認識精度の向上が必要である。

また、話者認識システムの実用化においては、認識精度の向上だけでなく、音声の継時変化に対処するための登録パターンの自動更新、風邪等で本人が棄却された場合の代替手順などのユーザに対する配慮が重要である。

参考文献

1.[Atal 72] B.S.Atal, Automatic speaker recognition based on pitch contours, J. Acoust. Soc. Am., 52, pp.1687-1697, 1972.

2.[Atal 74] B.S.Atal, Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, J. Acoust. Soc. Am., 55, pp. 1304-1312, 1974.

3.[Atal 76] B.S.Atal, Automatic Recognition of Speakers from Their Voices, Proc.IEEE, vol.6, no.4, pp.460-475, 1976.

4.[Doddington 85] G.R.Doddington, Speaker Recognition - Identifying People by their Voices, Proc. IEEE, vol.73, no.11, pp.1651-1664, 1985.

5.[Everett 85] S.Everett, Automatic speaker recognition using vocoded speech, Proc. ICASSP 85, pp.383-386, 1985.

6.[古井 74] 古井、音声の個人性パラメータの時期的変動と話者認識、信学論、Vol.J 57-A, No.12, pp.880-887, 1974.

7.[古井 77] 古井、話者認識における登録話者数の効果、信学会情報全大、p.228, 1977.

8.[Furui 81] S.Furui, Cepstral Analysis Technique for Automatic Speaker Verification, IEEE. Trans. ASSP, Vol.29, pp.254-272, 1981.

9.[古井 85] 古井、デジタル音声処理、東海大学出版会, pp.193-206, 1985.

10.[古井 91] 古井、話者識別技術；計測と制御、vol.25,No.8,pp.688-693, 1986.

11.[Furui 90] S.Furui, Speaker-Dependent-Feature Extraction. Recognition and Processing Techniques, Proc.Tutorial and Research Workshop on Speaker Characterization in Speech Technology, Edinburgh, Scottland, pp.10-27, 1990.

12.[Glenn 68] J.W.Glenn and N.Kleiner, Speaker Identification based on nasal phonation, J.Acoust. Soc. Am., 43-2, 1968.

13.[Hattori 92] H.Hattori, Text-independent speaker recognition using neural networks, Proc. ICASSP 92, II, pp.153-156, 1992.

14.[Hecker 71] M.H.L.Hecker, Speaker Recognition - An Interpretive Survey of the Literature, American Speech and Hearing Association, Washington, D.C., 1971.

15.[市川 79] 市川、中島、中田、電話音声を対象とした話者照合、日本音響学会誌、36, 2, pp.63-69, 1979.

16.[Johnson 32] C.Johnson, H.Hollien, and J.Hicks, Speaker identification utilizing selected temporal speech features, J. Phonetics, 12, pp.319-326, 1984.

17.[Li 83] K.P.Li and E.H.Wrench Jr., An approach to text-independent speaker

recognition with short utterances, Proc. ICASSP 83, pp.555-558, 1983.

18. [Luck 69] J.E.Luck, Automatic speaker verification using cepstral measurements, J.Acoust.Soc. Am., 46-4, 1969.

19. [Markel 79] J.D.Markel and B.Davis, Text-independent speaker recognition from a large linguistically unconstrained time spaced data base, IEEE Trans. ASSP, vol.ASSP-27, no.1, pp.74-82, 1979.

20. [Matsui 90] T.Matsui and S.Furui, Text-Independent Speaker Recognition Using Vocal Tract and Pitch Information, Trans.ICSLP 90, Kobe, 5.3, 1990.

21. [松井 91] 松井、古井、VQ、離散／連続HMMによるテキスト独立形話者認識法の比較検討、音声研資 SP91-89, pp.65-70,1991.

22. [Matsumoto 89] H.Matsumoto, Texi-independent speaker identification from short utterances based on piecewise discriminant analysis, Computer Speech and Language 3, pp.133-150, 1989.

23. [新美 79] 新美、音声認識、共立出版、pp.206-237, 1979.

24. [Naik 89] J.M.Naik, L.P.Netsch and G.R.Doddington, Speaker Verification over Long Distance Telephone Lines, Proc.ICASSP89, S10b.3,1989.

25. [Naik 90] J.M.Naik, Speaker Verification: A Tutorial, IEEE Communicatons Magazine, January, Vol28, No.1, pp.42-48, 1990.

26. [Openshaw 92] J.P.Openshaw, Z.P.Sun and J.S.Mason, A comparison of feature performance under degraded speech in speaker recgnition, Proc. ETRW - Speech Processing in Adverse Conditions, pp.119-122, 1992.

27. [O'Shaughnessy 86] D.O'Shaughnessy, Speaker Recognition, IEEE ASSP Magazine, pp.4-17, Oct. 1986.

28. [Pfeifer 78] L.Pfeifer, New techniques for text-independent speaker identicaon, Proc. ICASSP 78, pp.283-286, 1978.

29. [Poritz 82] A.B.Poritz, Linear Prodictive Hidden Markov Models, Proc. ICASSP 82, S11.5, 1982.

30. [Rose 90] R.C.Rose and D.A. Reynolds, Text independent speaker identification using automatic acoustic segmentation, Proc. ICASSP-90, pp.293-296, 1990.

31. [Rosenbarg 76] A.E.Rosenbarg, Automatic speaker verification, Proc. IEEE, 64, pp.475-478, 1976.

32. [Rosenberg 90] A.E.Rosenberg, C.H.Lee and F.K.Soong, Subword Unit Talker Verification Using Hidden Markov Models, Proc. ICASSP 90, S5.3, 1990.

33. [Rudasi 91] L.Rudasi et al., Text-independent talker identification with neural networks, Proc. ICASSP 91, pp.389-392, 1991.

34. [Sambur 75] M.Sambur, Selection of acoustic features for speaker identification, IEE Trans. ASSP, ASSP-23, pp.176-182, 1975.

35. [Savic 90] M.Savic and S.K.Gupta, Variable Parameter Speaker Verification System Based on Hidden Markov Modeling, Proc.ICASSP 90, S5.7, 1990.

36. [Schwartz 82] R.Schwartz, S.Roucos, and M.Berouti, The application of probability density estimation of text-indepndent speaker identification, Proc. ICASSP 82, pp.1649-1652, 1982.

37. [Sebestyen 62] G.Sebestyen, Pattern recognition by ana adaptive process of sample set construction, IRE Trans. IT-8, S82-S91. 1962.

38. [Shridhar 82] M.Shridhar and N.Mohankrishnan, Text-independent speaker recognition: A review and some new results, Speech Commun., vol.1, pp.257-267, 1982.

39. [白井 87] 白井、間野、石毛、ベクトル量子化スペクトルの頻度分布による話者認識、信学論 Vol.J 70-D, No.6, pp.1181-1188, 1987.

40. [Soong 85] F.K.Soong, A.E.Rosenbarg, L.R.Rabiner, and B.H.Juang, A vector quantization approach to speaker recognition, Proc. ICASSP 85, pp.387-390, 1985.

41. [Tseng 92] B.L.Tseng, F.K.Soong, and A.E.Rosenberg, Continuous probabilistic Acoustic Map for speaker recognition, Proc. ICASSP 92, II, pp.161-164, 1992

42. [Wolf 72] J.Wolf, Efficient acoustic parameters for speaker recognition, J. Acoust. Soc. Am., 51, pp.2044-2056, 1972.

43. [Yegnanarayana 92] B.Yegnanarayana, A.S.Madhukumar and V.R. Ramachandran, Robust features for applications in speech and speaker recogtion, Proc. ETRW - Speech Processing in Adverse Conditions, pp.97-100, 1992.

44. [Zheng 88] Y.-C.Zheng and B.-Z.Yuan, Text-Dependent Speaker Identification Using Circular Hidden Markov Models, Proc.ICASSP88,S13.3, 1988.

2.4 音声規則合成

2.4.1 はじめに

音声規則合成とは、少数の特徴パラメータとそれらを制御するルールによって音声を合成する方式であるが、最近では文字で書かれたテキストから音声に変換するテキスト・音声変換技術としてとらえられている。留守番電話に見られるＡＤＰＣＭ（Adaptive Differential Pulse Code Moduletion）等が、元になる音声を圧縮して符号化しておき、再生する録音再生方式であることに対して、音声規則合成は文字などの記号列を入力とした音声への変換装置である。これはちょうど、音楽の世界におけるミュージックシンセサイザにたとえられる。ミュージックシンセサイザがテンポや音の高さを変えられるように、規則合成でも発声速度や声の高さを変えることができる。ミュージックシンセサイザが音符を入力するのに対して、規則合成の入力は、英語や日本語などの文字である。そのため両者とも少ないデータ量で音が合成できる。しかし、両者とも人工的に音を作る制約から、生の音に比べると自然性や明瞭性が落ちてしまうのが現状である。

規則合成は、現在、米国やヨーロッパ、日本など多くの国で種々の言語について研究されている。その中でも日本語は、以下のように言語特有の難しさを持っている。

- べた書きされるので単語の区切りが明確でない。
- 同じ単語を漢字で表記したり、ひらがな、およびその混ぜ書きで表記したりするなど明確な正書法がない。他にも送りがなや数詞などの表記にゆれがある。
- 造語力が強く、漢字を組み合わせた新しい語が自由に作り得る。
- 丁寧語、尊敬語、謙譲語、古語など口語文法とは異なる文法体系に属する語が混在して使われる。

このように、規則合成は、言語に依存した部分が大きく、各言語に統一的な手法をとることは困難である。そのため、ここでは日本語の規則合成についての手法を紹介する。

2.4.2 形態素解析

形態素とは、それ以上分割できない語の単位である。形態素解析は、単語の切り出しや品詞の同定、未知語（辞書にない語）の処理を行う。形態素解析の目的は、文中の単語を正しく認定することである。さらに、係り受け解析のための前処理という位置づけもある。単語が認定されれば、その単語の読みやアクセントに関する情報が辞書から引き出せる。

日本語の場合、文節は自立語＋付属語列という単純な形式になっている。さらに、形態素と形態素の間には、接続の規則がある。例えば、「文書／を／処理／する。」では、格助詞の前

は名詞、動詞の前には助詞、サ変語尾の前にはサ変動詞に結合できるという規則により処理される。このような規則をマトリクス形式の接続表にし解析を進めていく。また、形態素を切り出す際、切り出し候補が多数存在する。切り出し位置が確定しても、その品詞にも複数の候補が存在する。したがって、形態素解析では、こうした多数の組み合わせから、解釈可能な候補を絞り込まなくてはならない。

そのための手法の１つとして、最長一致法がある。最長一致法は、与えられた文字列に対して、辞書から最も長い単語を優先する方式である。高速ではあるが、その根拠は明確ではないうえに、解析結果に尤度による優先順位を付けることができない［吉村82］。

文節数最小法は、文から語を切り出す際、文節の数が最も少ない候補を優先する方式である［吉村82］。この方式では１つの入力文に対して文節構造規則のみを満たす構造が複数個存在する場合には、文節数が少ない構造のほうが多いものより係り受け構造規則を満たす可能性が大きいはずであるという考えによっている。

この他にも動的計画法を用いて全ての単語候補から能率良く分割パタンを得る方法もある［藤崎81］。

２.４.３　係り受け解析

音声を生成するときには、ピッチパターン（基本周波数の変動の時間パターン）を生成する必要がある。ピッチパターンには、文節単位で変動する話調成分と音節単位で変動するアクセント成分からなる。話調成分は、２つの文節間で結合が強い場合は後ろの文節の話調成分が抑えられ、結合が弱い場合には後続文節の話調成分のレベルが大きくなる傾向がある［箱田80］。このような音声上の性質を持つために、文節間の結合の強さを決定する必要があり、このために係り受け解析を行うわけである。

文節間の結合の強さは形態素解析で得られた品詞や助詞の種類で決める方法もあるが、より自然なイントネーションを得るためには、文を句構造に分解し、文節間の結合度を決定する。句構造を得るには、構文解析（統語解析）といわれる自然言語処理技術が必要となる。

文法の形式の分類は、チョムスキーの階層として

・０型文法（一般文法）
・１型文法（文脈依存文法）
・２型文法（文脈自由文法）
・３型文法（正規文法）

の４つの階層が定義されている。

自然言語には、文脈自由文法に拡張を加えた文法がよく用いられている。文脈自由文法は句という単位を認定し、句のどのような組み合わせが、より大きな句になるかを書換規則の形で表したものである。プログラミング言語は文脈自由文法に制約を加えたものが多い。構文を解析するために、自然言語の構文上の曖昧さをうまく扱い、かつ効率のよい方法が工夫されている。その方法は、解の探索のやり方により、トップダウンかボトムアップ、縦型か横型という2つの組み合わせに分類される。トップダウン解析法は、解析の目標である文から始まり、それを文法規則にしたがって語に到達するまで解析していく。ボトムアップは、語から始まり、それを文法規則にしたがって文に到達するまで解析していく方法である。トップダウンでは、適用する規則の無駄が少ないが、無限に解を探索する危険がある。一方ボトムアップの場合には、解があれば確実に解にたどりつく。横型と縦型は、解の探索の方向に違いがあり、縦型では、1つの解について優先的に求めるが、横型では、解析は並列に展開されていく。縦型はある1つの解だけを求めるのに適し、横型は、全ての解を求めるのに適する。CKY法[Aho72]、チャートパーザ法［Kay80］は横型解析法の1つである。

2.4.4 アクセント処理

英語は強弱のアクセントであるが、日本語は高低のアクセントである。東京語においては、単語アクセントの種類は十数種程度と少ない。東京の単語アクセントの特徴の概略を述べると次のようになる。

- アクセントの高い拍が2カ所に別れて存在することがない。ただし「呉越同舟」のように複数の単語が組合わさって成句になっているものは除く。
- 第1拍と第2拍は高さが必ず違う。
- 最後が高いアクセントで終わるときは次の語がきた場合、高くつく場合と低くつく場合がある。

例えば、3拍の単語では、／低高高＋高／、／高低低／、／低高低／、／低高高＋低／の4種類のアクセント型があり、前から順番に、0型、1型、2型、3型と呼ばれている。0からnの数字は、アクセントの高いところから低いところに移る拍の番号を示しており、アクセント核といわれるものである。0型だけはアクセント核がないので、特別に数字の0を割り当てている。アクセント核がない0型を平板型ともいう。一般にn拍の単語の場合は、n＋1個のアクセント型が存在する。

これまで述べた説明は自立語といわれる語についてのものである。自立語は、これだけで文節を構成できるという性質をもっている。辞書にアクセントが載っている場合があるが、これ

も自立語になり得る単語に対してしか記述されていない。助詞などの付属語のアクセントは前接する自立語に大きく影響する。この規則は付属語アクセント結合として研究されている［匂坂83］。

自立語と付属語のアクセント結合は、5種類のパタンがあり、その結合の型は、それぞれ従属型、不完全支配型、融合型、支配型、平板型といわれているものがある。従属型は、結合後のアクセント型が自立語のアクセント型になる。不完全支配型は、自立語が平板型のとき、付属語のアクセント核が有効になる型である。融合型は、自立語が平板型でないとき付属語のアクセント核が有効になる型である。支配型は、いつでも付属語のアクセント核が有効になる型である。平板型は、常に0型になるもので実際には例が少ない。

従属型

捜す（0型）＋た　→　捜した（0型）

泳ぐ（2型）＋た　→　泳いだ（2型）

不完全支配型

捜す（0型）＋たり（1型）　→　捜したり（3＋1→4型）

泳ぐ（2型）＋たり（1型）　→　泳いだり（2型）

融合型

捜す（0型）＋ながら（1型）　→　捜しながら（0型）

泳ぐ（2型）＋ながら（1型）　→　泳ぎながら（3＋1→4型）

支配型

捜す（0型）＋まい（1型）　→　捜すまい（3＋1→4型）

泳ぐ（2型）＋まい（1型）　→　泳ぐまい（3＋1→4型）

平板型

捜す（0型）＋だけ　→　捜すだけ（0型）

泳ぐ（2型）＋たけ　→　泳ぐだけ（0型）

複合単語の場合もアクセント結合が生じる。後続単語が有核型のアクセントの場合は、付属語のアクセント結合様式の支配型に相当する自立語結合保存型が普通である。後続単語が平板型、尾高型の名詞である場合は、後続単語の先頭のモーラにアクセント核が移動する（自立語

結合生起型）。さらに後続単語が２モーラ以下の場合は、前の単語の一番後ろにアクセント核が移動する接辞結合標準型の場合と、全体が平板化される接辞結合平板化型に分けられる。

自立語結合保存型

予算（１型）＋市議会（２型）　→　予算市議会（３＋２→５型）

自立語結合生起型

球技（１型）＋大会（０型）　→　球技大会（３＋１→４型）

接辞結合標準型

美術（１型）＋展（０型）　→　美術展（３＋０→３型）

接辞結合平板化型

美術（１型）＋的（０型）　→　美術的（０型）

２.４.５　音韻処理

日本語では、仮名表記が、ほぼそのまま音声に対応する。しかし以下のように音韻に対する処理が必要な場合がある。

（１）助詞「は」「へ」「を」の場合

この場合は、「は」→「ワ」、「へ」→「エ」、「を」→「オ」のように変換する。「こんにちは」のように、形態素解析で助詞として判断されない場合もあるので、このような語は音声表記を辞書に持っておく必要がある。

（２）長音化

映画（えいが→エーガ）、経営（けいえい→ケーエー）、氷（こおり→コーリ）などのように「オウ」、「エイ」、や同一母音の連続は長音化される。しかし「姪」などの和語や「小売」のように漢字境界にまたがるものなどは長音化されない。

（３）鼻音化

ガ行音は、語頭以外は原則として鼻音化する。ただし、「ゴトゴト」のように擬声語で同じ音が繰り返されるとき、数詞の「五」、接頭語の次のガ行等は鼻音化しない。

（4）無声化

本来有声音である音が、口の構えだけ残して声帯を振動させないで発声することがあるが、これを母音の無声化という。「キ」「ク」「シ」「ス」「チ」「ツ」「ヒ」「フ」「ピ」「プ」「シュ」などの拍が、「カ」「サ」「タ」「ハ」「パ」の直前にきた場合、無声化する。しかし、アクセント核がある音節は無声化しない。

（5）連濁

「カブシキ」（株式）と「カイシャ」（会社）が結合すると「カブシキガイシャ」とカイシャの先頭音節が濁音化する。この現象を連濁という。連濁は、必ずおこる現象でもない。連濁現象は極めて複雑であるが、基本的な性質を次に揚げる［佐藤83］。

- 複合語第２語の第２音節以降に濁音が存在するときは連濁しない
- 第２語が漢語のときは連濁しない
- 複合動詞は連濁しない
- 擬声、擬態語は連濁を起こさないが、名詞や動詞の畳語は連濁する
- 音韻レベルで同じ音韻が続かないように連濁することがある
- 連濁を起こしにくい接頭語がある
- 複合語が右枝分かれ構造の場合連濁しにくい

（6）数詞

日本語における数詞の表記には、算用数字、漢数字が用いられ、その表記が多様であるばかりでなく数詞に助数詞が連接することにより、数詞や助数詞の音韻やアクセントの結合が起こるなど複雑である。［宮崎84］では、数詞や助数詞の音韻変化に着目した助数詞の分類や音韻とアクセントの変化表によって数詞処理を行う方法を提案している。

2.4.6　音声素片

規則合成における音声の最小の合成単位が音声素片である。入力されたテキストは、この音声素片の単位に分解され、各音声素片を接続して合成音声を得る。音声生成の元になる音声素片は、各言語によって最適なものが検討されている。

一番少ない素片数で構成するならば、音声素片は音素がよく30～50個程度で実現できるが、音素と音素を結合する方法が確立されてないため、良い品質の合成音が得られていない。日本語を構成する最小音節がＣＶ（子音－母音）で表されるため、これを素片に使うことも多い。ＣＶだけだと母音から子音への遷移で滑らかさに欠けることがあるためＶＣ素片を追加するこ

ともある。近年半導体技術が進歩してきてメモリが大容量になってきた背景もあり、最近はＶＣＶ［佐藤78］やＣＶＣ［佐藤84］、さらには大量のデータベースを用いて、より長い素片を使う［匂坂87］ことも多くなった。長い素片を使えば、滑らかさは向上するが、一つ一つの音韻のバランスを制御することが難しくなってくる。

子音－母音の組み合わせだけでなく、調音結合による音韻変形の反映した合成単位を、統計的手法により自動的に生成する規則合成法についても検討されている。ＣＯＣ法［中嶌88］では、音韻ラベル付けされた学習用音声データを用い、「音素環境に基づくクラスタリング法」（Context Oriented Clustering：COC 法）と呼ぶ方法によって、調音結合による変形の著しい音韻（広義の異音）を探索し、その音韻環境情報（先行・後続音韻）と共に合成単位として蓄える。

２.４.７　音韻時間長制御

日本語音声にはモーラという音節に対応する感覚的な時間単位が存在するが、自然音声の音韻の時間長を調べてみると決して一定ではない。音韻時間長がどんな要因によって決定されているか発話データの時間長計測に基づいて定式化した研究がある［匂坂81］。これによれば、子音長は、呼気段落内のモーラ数と子音の性質に依存した固定量によって決定する。また母音時間長は隣接する子音の影響を受ける。

日本語では、音韻を拍という概念で数えることから、拍同期点という拍の基準点を音韻ごとに設け、それを等間隔に並べる方法もある［阿部81］。拍同期点は、すべてのＣＶ音節について、[maＣＶma]を発声し、音節[ma]の母音開始点を基準にして相対的に位置を求める。

さらに、後続音節に依存して受聴タイミング点を決め音節を不等間隔に配置する方法もある［Minowa 86］。この方法も拍同期点の場合と同様に、すべてのＣＶについて受聴タイミング点を定める。[aＣＶa]の組み合わせの試験音声を作り、聴感試験によって３つの音節が等間隔に聞こえるものを選び、[a]と[a]の立ち上がり時点の中間に位置するＣＶ上の位置をその音節の受聴タイミング点とする。この受聴タイミング点を基準点として、後続音節によって定まる受聴タイミング点間隔を用いて、順次ＣＶ音節を配置する。後続音節との受聴タイミング点間隔をＬ、平均間隔をＬtとするとＣＶ音節配置規則は

$$L = Ap \cdot As \cdot Ab \cdot Lt$$

のように表される。ここでＡbは後続音節によって定まる受聴タイミング点間隔をＬtで正規化した係数、Ａsは撥音、促音、長音、連母音、無声化音に対する補正係数、Ａpは文頭音節に対する短縮係数である。［荒居89］

２.４.８　ピッチの生成

標準語では、一息で発声される区間のピッチ周波数は、時間とともに降下する。この成分は話調成分といわれ、これにアクセントによるピッチの時間パタンが重畳される。現在ピッチパタンの生成方法として、点ピッチ法［橋本73］と臨界制動モデル［藤崎71］の２つがよく使われる。

点ピッチ法は、音韻の母音の重心点におけるピッチ周波数がきわめて安定であることを利用し、重心点を直線で接続していく。

臨界制動モデルでは、ピッチパタンは、大脳言語中枢からの種々の指令が、神経系を経て声門下圧、声帯緊張を支配する筋肉系に作用し声帯発振を起こす様子をモデル化したものである。声立ておよびアクセントに関する指令は単位ステップ関数とし、前者は発声開始の直前から停止の直前まで連続的に加えられ、後者は単語のアクセント型に対応して特定のモーラの発音とほぼ同期して加えられる。それぞれの制御機構の特性はともにステップ応答で記述されるものとし、それぞれ　Gv(t)、Ga(t)と表現する。声立てのステップ応答 Gv(t) はいったん上昇した後、次第に下降する時間波形であり、アクセントのステップ応答Ga(t)はt=0で連続な微係数を有する単調非減少の波形である。これらの特徴を有する関数のうち、最も簡単なものは、

$$Gv(t)=Av\cdot\alpha\cdot t\cdot\exp(-\alpha\cdot t)$$

$$Ga(t)=Aa\cdot[1-(1+\beta\cdot t)\exp(-\beta\cdot t)]$$

で表せる。ここでAv、Aaはそれぞれの応答の大きさを表す振幅パラメータ、α、βは応答速度を表す時間パラメータである。図２．４-１にこれらの関数の時間波形を示す。

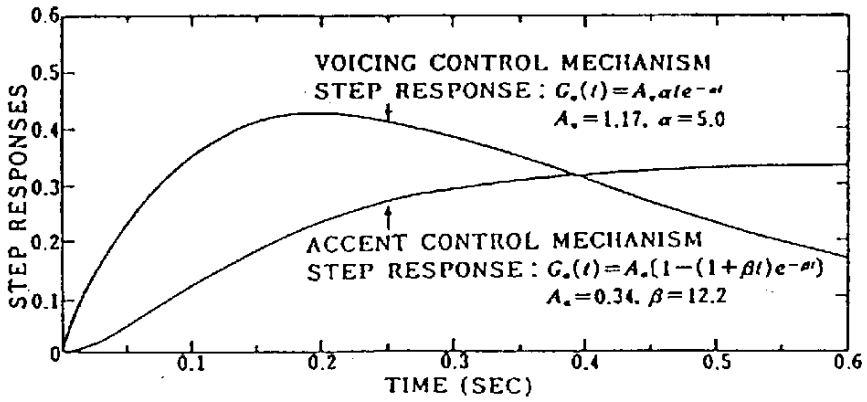


図２．４-１　ステップ応答時間波形（［藤崎71］による）

図２.４-２は、この応答関数を使ったピッチ生成モデルで、フレーズ指令のステップ信号とアクセントステップ信号に対する２つの応答関数の和でピッチを表している。

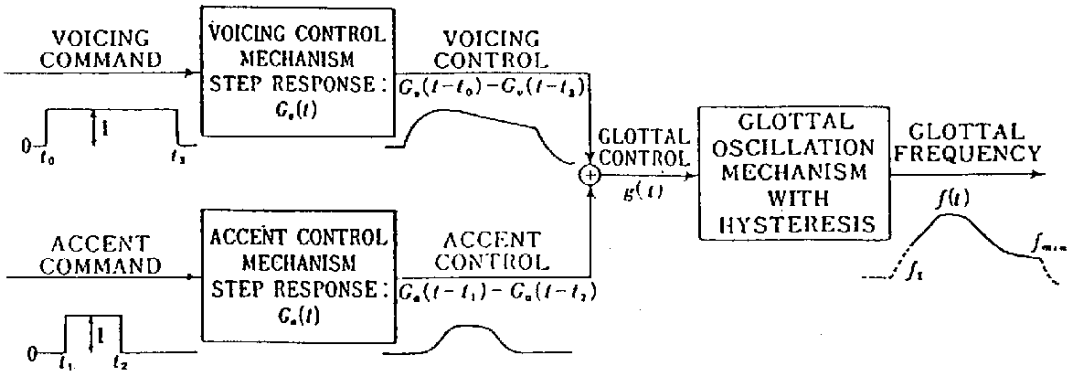


図２.４-２　ピッチ生成モデル（［藤崎71］による）

２.４.９　合成方式

素片が接続されピッチパタンが用意できれば、それらを合成部で処理すると音声波形が得られる。その合成部には幾つかの方式がある。ここでは、合成方式について述べるが、音声合成の方式と素片を表すパラメータの導出法には深い関係がある。したがって合成方式の優劣はパラメータの導出法も含めた総合的な比較で決まることに注意したい。

音声の特徴を表すものの１つにホルマントがある。特に米国では、ホルマント合成方式がさかんに検討され実用の域に達している［Klatt 80］。スペクトルパラメータはホルマント周波数と帯域幅で制御する。合成器が比較的容易にできるが、規則合成器に適したパラメータを得るのが難しい。

ＬＳＰ方式は、線形予測系の技術から生まれたものであり、ＰＡＲＣＯＲ方式とも関係が深い。ＰＡＲＣＯＲ係数が時間領域のパラメータであるのに対し、ＬＳＰパラメータは周波数領域のパラメータであるため、粗く標本化して、その間を直線補間しても歪が小さいという長所がある［板倉 75］。合成器の規模は中程度である。

声道アナログ方式は、声道断面積をパラメータとした方式で、人間の発声機構をモデルにされたものである［Flanagan 80］。男性や女性の差や、個人性などもパラメータに反映しやすく理想的ではあるが、規則合成に適したパラメータを得る手法が確立しておらず、良好な音声

が得るのが難しい。

ケプストラム方式は、ケプストラム係数をパラメータにして合成する方式である。線形予測系のスペクトルモデルが全極型であるのに対し、ケプストラム方式は対数スペクトルの形状がそのままパラメータになるので極の数等を考えなくても良い。さらに周波数軸にメル尺度を採用したメルケプストラム方式もある［古市 84］。利点としては、規則合成に適したパラメータが得やすいことであるが、合成器がやや複雑である。

素片編集方式は、スペクトルパラメータではなく1ピッチの波形を重ね合わせるもので合成が簡単である。

2.4.10今後の動向

現在の規則合成は、形態素解析、係り受け解析が行われている状況であるが、このレベルでは同形語の判別や、強調表現は難しい。今後、精度向上を行うには意味解析まで含んだ、より深い解析が必要になる。同形語について例をあげれば、「大家」は「オオヤ」という読み方と「タイカ」という読み方があるが、両方とも名詞であり文法的に選別することはできない。他にも、「日」は、同じ「ヒ」という読み方でも、アクセントが0型と1型では意味が違っているような例もある。強調の例では、「彼は妹の友達である。」という文の場合、「彼は／妹の友達である。」のように文を切ると妹が強調されるが、「彼は妹の／友達である。」のように切ると友達が強調される。どちらが正しいかは、文脈による。言語的に強調するところがわかっても、物理的なピッチパタンで表すことができなければ音声に変換できない。これについては［武田 91］［白井 87］などの研究がある。

さらに書き言葉だけでなく会話文も処理できるような処理系が要求されてくる。このために辞書や文法の拡張が行われていくだろう。

また、感情表現が自由に制御できるようになると、音声の特徴を生かした用途が広がってくると考えられる。現在この方面での研究も始まっていて、怒り、歓喜、悲哀を表現するための韻律制御ルールが検討されている［北原 89］。これによると怒りは、パワーの増大に伴って各文節の高ピッチ部の振幅が増大、また、文尾を除く文章全体のピッチのダイナミックレンジが増大する。歓喜は、ピッチのダイナミックレンジが増大する。悲哀は、ピッチのダイナミックレンジが小さくなり、発声時間が長くなるという傾向が観察されている。

参考文献

1.［吉村 82］吉村賢治、日高達、吉田将、"最長一致法と文節数最小法について"、情報

処理学会論文誌、23、6

2. [吉村 82] 吉村賢治、日高達、吉田将、"文節数最小法を用いたべた書き日本語文の形態素解析"、情報処理学会論文誌、23、6

3. [藤崎 81] 藤崎博也、"動的計画法による漢字仮名混じり文の単位切りと仮名ふり"、情報処理学会自然言語処理研究会資料 28-5

4. [野口 91] 野口正一、牧野武則、自然言語処理、オーム社

5. [Aho 72] Aho,A.V.and Ulman,J.D.、The Theory of Parsing,Translating, and Compling、Vol.1、Prentice-Hall

6. [Kay 80] Kay,M、"Algorithm Schema and Data Structure in Syntactic Processing"、CSL-80-12、Xerox Palo Alto Research Center

7. [Earley 70] Earley,J.、"An Efficient Context-Free Parsing Algorithm"、CACM、Vol.13、No.2、pp.94-102

9. [箱田 80] 箱田和雄、佐藤大和、"文音声合成における音調規則"、信学論、J63-D、No.9、pp.715-722

10. [匂坂 83] 匂坂芳典、佐藤大和、"日本語単語連鎖のアクセント規則"、電子通信学会論文誌、Vol.J66-D、No.7、pp.849-856,1983

11. [佐藤 78] 佐藤大和、"ＰＡＲＣＯＲ－ＶＣＶ連鎖を用いた音声合成方式"、信学論、J61-D、11、pp.858-865

12. [佐藤 84] 佐藤大和、"ＣＶＣと音源要素に基づく（ＳＹＭＰＬＥ）音声合成"、音響学会音声研究会資料、S83-69

13. [匂坂 87] 匂坂芳典、"種々の複合音声単位からの音声合成"、音響学会講演論文集、1-6-9、pp.173-174（1987秋）

14. [中嶌 88] 中嶌信弥、箱田和雄、"音韻環境に基づくクラスタリングによる規則合成法の評価"、音響学会講演論文集、1-2-20、pp.139-140（1988秋）

15. [宮崎 84] 宮崎政弘、"日本文音声変換のための数詞読み規則"、情報処理、Vol25、No.6、pp.1035-1043

16. [佐藤 83] 佐藤大和、"連濁の分析と規則化の検討"、音響学会講演論文集、1-2-10、pp.61-62（1983秋）

17. [阿部 81] 阿部芳春、今井聖、"音声合成における音節の時間配置の法則化"、音響学会講演論文集、3-4-8、pp.633-634

18. [匂坂 81] 匂坂芳典、東倉洋一、"音韻固有の性質を考慮した音韻継続時間長設定"、音響学会音声研究会資料、S80-72

19. [匂坂 84] 匂坂芳典、東倉洋一、"規則による音声合成のための音韻時間長制御"、信

学論、J67-A、7、pp.629-638

20. [Minowa 86] Minowa,T. and Arai,Y.、"The Japanese CV-syllable positioning rule for speech synthesis"

21. [藤崎 71] 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ンとその生成機構のモデル"、音響学会誌、27、9、pp.445-453

22. [橋本 73] 橋本新一郎、"日本語アクセントの諸性質"、信学論、J56-D、11、pp.654-661

23. [箱田 80] 箱田和雄、佐藤大和、"文音声合成における音調規則"、信学論、J63-D、9、pp.715-722

24. [Flanagan 80] Flanagan,J.L.、Ishizaka,K.、Shipley,K.L.、"Signal models for low bitrate coding of speech"、JASA、68、3、pp.780-791

25. [Klatt 80] Klatt,D.H、"Software for a cascade/parrallel formant synthesizer"、J.Acoust.Soc.Amer.、67、3、pp.971-995

26. [板倉 75] 板倉文忠、"線形予測係数の線スペクトル表現"、音響学会音声研究会資料、S75-34

29. [古市 84] 古市千枝子、今井聖、"CV音節のメルケプストラムパラメータの接続に基づく音声の規則合成"、信学論、J61-D、11、pp.1356-1363

30. [北原 89] 北原義典、東倉洋一、"音声の韻律成分制御による感情表現"、音響学会講演論文集、2-7-12,pp.193

31. [武田 91] 武田昌一、市川薫、"日本語文音声のプロミネンス生成規則の作成と評価"、日本音響学会会誌47、6

32. [白井 87] 白井克彦、岩田和彦、"音声合成のための単語の強調表現の規則化"、電子情報通信学会論文A、J70-A、pp.816-821

33. [荒居 89] 荒居康彦、大崎正巳、音声処理とDSP、啓学出版

第３章　応用システム別の技術課題

第3章　応用システム別の技術課題

3.1　音声ワープロ

3.1.1背　景

情報処理システムの進展と日本語ワープロの普及によって、現在では大半の文書がワープロによって作成されている。今後もこの傾向が進むことには疑いがなく、文書入力作業の一層の負担軽減は今後ますます強く望まれることになろう。これまでのワープロの文書入力手段のほとんどはキーボードであり、最近になってペン入力も使われるようになりつつある。これらの入力手段に比べて、音声による文書入力は、入力時にも手、視線、姿勢が自由になるため他の作業を行ないながらの文書入力が可能、人が日常的に使用しているコミュニケーション手段であり万人にとって抵抗感が少ない、原理的な情報伝達速度が大きく高い入力効率が期待できるなどの利点を持ち、その実現が望まれる。

ここで述べる音声ワープロは、作成すべき文書内容の一言一句を音声で入力するものに限らず、文書作成を目的として音声認識·理解を行なうシステムを広く含むものとする。以下では、まず音声ワープロの使用形態、技術要素と研究課題について述べたのち、現在の音声ワープロの研究例を紹介する。

3.1.2　音声ワードプロセッサの使用形態、利点、必要度

（1）特定業務文書作成

作業報告書、検査報告書、医療診断書など特定の業務で使用される文書を作成する。検査作業、診断業務等を行ないながら文書を作成することが多く、手、視線、姿勢が自由になるという音声入力の特徴のためにユーザが作業、思考を中断することなく文章を作成することが可能になる。医療診断書の場合には、現在は医師が口述した録音テープを人が書き起こしているが、これに比べて書き起こしのコストを低減できるうえ、診断書をすぐに手にすることができ、内容の確認や訂正も容易になる。

この使用形態では、文書内容、用語、文章表現、様式が決められていることが多いことから、音声認識の対象となる語彙、文法をある程度限定することが可能であり、実現可能性が高い。キーワードなど特定の単語を音声入力することによって定型文を呼び出し、その定型文中の語句を埋めることによって文書を高速に作成することが可能である。ユーザが特定されるため話者適応を行なうことができ、また発声の仕方を訓練することも可能である。雑音環境は業務内容に依存する。

（2）自由文書作成

現在のワープロの一般的な使用形態であり、任意の文書を作成するものである。音声入力によって、キーボードが苦手な人や身体障害者にも文書の作成が容易になる。キーボードの使用に慣れている人にとっては当面の間、誤認識訂正を含めた場合の音声入力の速度はキーボード入力よりも速いとはいえず、また多量の文書を音声入力する肉体的負担も大きいことから、キーボード入力のすべてが音声入力に置き換わるとは考えにくい。しかし、ペン入力に比べると入力速度が速く、また姿勢や動作にあまり影響されずに入力できることから、携帯情報機器等でのメモ入力などの少量の文書入力には有用である。

この使用形態では文書内容、用語、表現形態、様式等が自由であり、任意テキストの音声認識が必要となる。また一般にパーソナル機器での使用であるためユーザは特定可能であるが、使用頻度が低くあらかじめの話者適応が困難である場合もあり、不特定話者認識あるいは高速話者適応や使用しながらの話者適応が望まれる。またユーザの年齢層、経験、スキルなどが多岐にわたり、情報機器に不慣れなユーザも多いため、特に使いやすいユーザインタフェースが必要となる。発声法、使用法などに対するユーザの慣れは期待できる。使用環境はデスクトップ型、ラップトップ型ではオフィス、家庭など比較的静かな場所であるが、携帯型の場合は戸外で使用することも多い。

（3）速記メモ

インタビューや会議等の発言を音声認識し、速記メモを作成する。現在は速記者が行なっている速記作業を機械で置き換えるものであり、大きな省力化が期待できる。ただし、記録すべき音声はそもそも音声入力を意識して発せられたものではないため、言い誤り、言い直しなどを多く含み、語尾や文末が曖昧であることも多く認識は極めて困難である。発言者が感情的になる結果、発声が大声、早口になることもあり、これらへの対処も新たな研究課題である。しかも認識誤りに対してその場では再発声などによる修正を行なうことができない。その反面、速記メモからインタビュー記事や議事録を作成する際に録音を聞きながらの後修正が可能であるため、多少の認識誤りは許される。後修正の場面では、速記メモに対して録音された音声を自動的に位置付けるなどのユーザインタフェースによる支援が重要である。話者は不特定多数であり、装置への慣れもあまり期待できない。複数人の同時発言などもある。会議室等の室内での使用が主体となるため、周辺雑音は比較的コントロールしやすい。

（4）電話口述筆記

新聞記者等が取材先から記事を送る場合に、どこにでもある通信手段として電話を用いることが多い。記者はあらかじめ書いておいた原稿を読み上げるか、またはその場で原稿を口述筆

記する。現在は新聞社側で人が聞きとった上で改めて計算機システムへの入力作業を行なっているが、この聞き取り、原稿入力を音声認識を用いて自動化することによって、大幅な省力化、効率化を行なうことができる。記事内容は多岐に渡り認識対象語彙は多いが、業務上の音声入力であるため話者を特定でき、訓練による丁寧な発声を期待できる。ただし、電話回線を通じての音声認識であるため、周波数帯域が狭いこと、回線の伝送特性が変動すること、種々の音響特性を持つ電話機が使用されることなど、多くの認識性能低下の要因がある。また、入力手段が電話であるので、音声コマンドやプッシュ音を用いての使いやすいユーザインタフェースが必要である。出先での入力のため電話の設置場所で入力するため、まわりの話し声、車等の騒音に対する耐雑音認識が必要であることが多い。

以上のいずれの使用形態においても、実用化のためには単語認識率98%以上の認識性能の確保が必要であると思われる。

また、装置形態としては以下が考えられる。

(a) ワークステーション、パソコン、ワープロに音声入力機能があらかじめ内蔵、
(b) ワークステーション、パソコン、ワープロ用の増設ソフトウエアあるいは増設ハードウエア(ボード)、
(c) ワークステーション、パソコン、ワープロと接続して使用する単体音声入力装置、
(d) 携帯情報機器に内蔵、

いずれも音声ワープロは音声入力部(AD変換、分析)、音声認識部、文書編集部から構成される。各使用形態の特性の相違を表3.1-1にまとめる。

表3.1-1　音声ワープロの使用形態とその特性

	特定業務	自由文書	速記メモ	電話口述
語　彙	中～大語彙	大語彙	大語彙	大語彙
文法拘束	大	小	小	小
発声単位	単語/文節/連続	文節/連続	連　続	文節/連続
話　者	特　定	特　定	不特定	特　定
雑音環境	静か～雑音大	静　か	静　か	雑音大
UI	マルチモーダル	マルチモーダル	後編集	音　声
ユーザ訓練	可　能	場合により可能	不　可	可　能

3.1.3 技術要素と研究課題

（1）音素認識

音声ワープロによる文書作成のためには入力された音声の一音一音を正しく認識する必要があり、このためには音素、音節などの短い単位音声の高い認識性能が不可欠である。特に、音声ワープロで避けられない未知語の認識のためには、単語辞書等の言語知識を用いることができないため、単位音声の認識性能を一層向上させなければならない。

単位音声の認識のためには、個々の単位音声をHMMなどでモデル化した上でこれを多量の音声データで学習することによって単位音声のモデルを作る方法が一般的である。ここで認識精度を向上させるためには、単位音声が連続して発声されたときに隣接音素間の調音結合によって生ずる音声パタンの変形に対処しなければならない。すなわち、言語的には同じ記号で表現される音素であっても、単語や文中での音素環境によって音響的に異なる音声パターンとなる。これに対処するための方法としては、音素よりも小さい単位を認識の基本単位とすることによって、調音結合の影響を緩和する方法と、調音結合による変形部分を含んだパターンやモデルを認識の基本単位とする方法などがあり効果を上げている。

調音結合による変形への対処とともに、単位音声の基本的な識別性能の向上が必要であり、識別学習等によるモデル自身の識別能力の強化、音素モデル等の継続時間長制御による音素の挿入・脱落の誤り低減が必要である。また、単語辞書を用いることができない未知語の認識では、音素/音節/文字bigramやtrigramなどの統計言語情報の使用が有効である。

また、音素認識の性能向上のためには、単語認識において単語などの大きな単位を認識する場合に比べて、音声の時間方向の分析をより細かく行なう必要がある。このため、音声分析手法やフレーム長、フレーム周期等の分析パラメータを音素認識に適したものにチューニングすることが必要になる。

（2）大語彙単語認識

音声ワープロでは任意の文書内容の入力が要求されるため、語彙サイズが数千～数十万語の大語彙認識の技術が必要となる。すべての認識対象単語の音声をあらかじめ発声(登録)することは不可能であるので、かな文字列で定義された単語辞書を準備した上で、音素、音節等の単位音声の列として単語を認識する。例外として、特定業務文書作成の場合には語彙をある程度限定することが可能であるが、この場合でも1000語以上の語彙は必要になる。

大語彙単語認識の実現のためには、類似単語の増加による認識誤りに対処するために音素、音節等の単位音声の認識性能を向上させることのほかに認識のための計算量低減が必要になる。認識の計算量低減のためには、ビームサーチ、ベストファーストサーチなど探索のための計算

量削減、音声パターンのベクトル量子化とベクトルコード間の距離テーブルを用いた距離計算量削減が行なわれるが、これらの計算量削減に伴って最適解が保証できなくなることやモデルの精度、計算精度が低下することに起因して認識精度が一般に低下する。認識精度の低下を最小限に抑えながら、最大限の計算量削減を行なう方法の開発が課題となる。

同時に、認識精度を確保した上での計算量削減には限度があるため、予備選択技術が重要となる。計算量の少ない大まかな認識で入力音声に類似する単語を予備選択した上で、詳細識別を行なう方法である。正解単語が予備選択段階で落されないようにしなければならない。

（3）未知語入力

どんなに大語彙の単語辞書をあらかじめ準備しておいても、固有名詞等の未知語の出現は避けられず、これを入力する手段を備えていなければならない。このためには、まず入力音声中の未知語の部分を自動的に検出する手段が必要であり、さらに未知語を認識する手段、さらに今後の音声入力に備えて未知語を登録する手段が必要である。

未知語の検出、認識は単語辞書の拘束を与えない状態での音素列、音節列の認識を行う必要がある。この際、音素/音節/文字bigramやtrigramなどの統計言語情報による拘束を使用するのが有効である。

（4）連続音声認識

音声ワープロの入力単位はユーザが単語、文節等の文法単位を意識せずに発声できることが使いやすさの上で望ましい。特に日本語では単語や文節がテキスト上で連続しており、それらを正確に区切ることは一般には容易でない。このためには単語、単文節、連文節、文などの任意の単位で、途中にポーズを置かずに発声された連続音声の認識が必要である。ただし、特定業務文書作成における定型文入力の場合にはキーワード等の離散単語入力も考えられる。

現状の連続音声認識は、単語列に対して、有限状態オートマトンや文脈自由文法(CFG)などによる文法拘束や単語bigramやtrigramの統計言語情報による拘束をかけ、この拘束の元での文全体の尤度を最大化するような単語列を探索によって求める方法が一般的である。この探索には通常、ビタビ探索が用いられる。特定業務文書作成においてはそれぞれの業務文書に特化した比較的強い文法拘束をかけることでき、また語彙サイズが小さい場合もあるため、現状の方法である程度対応することは可能である。

しかしながら、自由文書作成、速記メモ、電話口述筆記の場合には認識対象が大語彙になるとともに、あまり強い文法拘束をかけることができないために、現状の方式では入力音声と照合すべき単語列が組合せ的に爆発し、計算量が大幅に増加するとともに、認識誤りが増加する。

強い文法拘束をかけられない場合には単語bigramやtrigramの統計言語情報による言語拘束

が有用であり、効果を上げている。しかしながら、精度の良い統計言語情報を得るためには非常に多量のテキストデータベースが必要となるほか、音声ワープロの入力文書と類似する分野、話題のテキストデータベースによって学習しなければ高い効果は期待できない。これに対しては、入力文書に類似する少量のテキストを用いて統計言語モデルを適応化する方法が検討されている。さらに、文書入力を行ないながら統計言語モデルを逐次的に自動的に入力テキストに適応させる技術が必要である。

音声ワープロにおいて入力された文書を修正する場合には、文あるいは文節の途中から音声入力されることがあり得る。したがって、文法の途中から始まる単語列を入力できる枠組が新たに必要となる。

（5）音声対話

音声ワープロの将来の姿として、現在の秘書による口述筆記のように、ユーザが文書の内容を比較的にラフに入力したものをもとに音声ワープロが文書を作成するという形態が考えられる。このためには、音声ワープロが作成している文書内容に関する専門知識や常識などを持ってユーザの発話を理解することが必要になる。また、文書を作成するために足りない情報をユーザに問い合わせるなどの音声対話インタフェースが必要となる。

特定業務文書作成における定型文中の特定の語句を埋めることによって文書を作成する形態で、上記の音声対話による文書作成の非常に初期の形態を実現することができる。この場合、ユーザは文書作成に必要なキーワード、専門用語を順に音声入力し、音声ワープロが入力された語句を用いて文書を完成させる。さらに、定型文中のまだ埋められていない部分や曖昧な文があれば、それを音声ワープロからユーザに問い合わせる。

（6）話者モデル学習・適応化

話者については不特定話者認識が可能であることが望ましいが、音声ワープロは速記メモを除いては一般的にパーソナル機器であるため、話者が限定できる場合が多い。この場合には話者適応をおこなうことによって大語彙および連続音声認識の性能を向上させるほうがよい。ただし、音声ワープロの使い始めの段階や少量の文書を入力する場合には、不特定話者認識によってもある程度動作することは大切である。さらに、音声ワープロの使用頻度の低いユーザにとって、話者適応のために毎回数十単語の発声を強いることはできないため、少数の単語の発声によって高速に話者適応をおこなったり、または文書入力を行ないながらの話者適応によって認識性能が徐々に向上する、いわゆるオンライン話者適応が望まれる。速記メモの場合も、発言者に話者モデル学習用の多量の音声の発声を強要することはできないため、同様のことが言える。

（7）言語モデル学習・適応化

音声ワープロではユーザによって音声入力されるテキスト内容は著しく異なり、さらには同じユーザでも毎回文書内容は大きく異なることは普通である。これらをすべて受け付ける汎用の言語モデルを用意することはできず、また汎用の言語モデルでは認識性能の向上はあまり望めない。認識誤りを低減し、入力効率を向上させるには、言語モデルをユーザの入力文書に対して高速に適応化することが必要である。特に、同じような認識誤りが何回も起こるとユーザにとっては精神的にも負担になるため、文書入力途中でのダイナミックな適応化が望まれる。

（8）耐雑音認識

特定業務文書作成では業務を行ないながら音声入力をおこなうことが多く、業務内容によってオフィス等の静かな場所から工場の高騒音下までのさまざまな周辺雑音環境で使用されることが考えられるため、雑音処理や接話型マイクなどを使用した耐雑音認識が必要である。この場合、一定の場所で使用することが多く、使用場所での雑音環境に適応できることが望ましい。自由入力の場合の多くはオフィスや家庭などの静かな環境での使用されるが、携帯機における音声入力の場合には戸外や車内等で使用が考えられるため、耐雑音認識が必要となる。速記メモ作成については、会議室等の閉じられた室内での使用が主体となるため、周辺雑音は比較的コントロールしやすいと思われ、通常のオフィス程度の雑音レベルに対応できれば良い。電話口述筆記の場合には出先からの音声入力となり、特に取材現場においては他人の電話などの話し声が雑音として入力されるため、この対処が必要である。また、屋外では車等の騒音にも対処しなければならない。

（9）電話音声認識

電話口述筆記では電話を通じた音声の認識をおこなう必要があるため、電話帯域の音声認識、回線特性の変動、種々の電話機の音響特性への対処が必要である。

（10）低コスト音声認識

現在普及しているワープロに替えて音声ワープロを広く普及させるために音声認識部分の低価格化が不可欠である。MPU、DSP等のハードウエアは今後も年々高速に、かつ低価格になっていくものと思われるが、依然として計算量低減が必要である。

（11）遠隔マイク入力

従来からよく使用されているヘッドセットマイクは圧迫感がある、席を離れられない、セットの手間がかかるなどの欠点があり、音声入力を気軽に行う上での障害となる。また、手持ち

マイクは手がふさがるという欠点がある。ユーザの負担軽減のためには、マイクが音声ワープロ本体に内蔵されるか据え置かれていて、マイクから数十センチ離れた位置から入力できるハンズフリーマイクの使用が望ましい。さらにユーザが姿勢を自由に変えることができるためには、マイクとの位置関係が変わっても認識性能に影響しないことが必要である。口とマイクとの距離が大きくなると周辺雑音の混入が多くなるほか、ユーザがその距離を一定に保つことが困難になるため入力音声レベルの変動が大きくなる。指向性の強いマイクや複数個のマイクを使用しての雑音除去やマイクロフォンアレイを用いて音声の発生源を追尾することなどが必要になろう。

（１２）話者識別

複数の発言者を発言を記録する速記メモの場合は、入力された音声の発言者を識別することが必要になる場合がある。この場合には各発言者にはあらかじめ少量の音声を発声してもらい、これから話者識別のための話者モデルを作成する方法がある。あるいは各話者にマイクを付けたり、単一指向性マイクを各話者に向けることによって、さらにはマイクロフォンアレイを用いて音声の発声源の位置を検出することによって話者識別をおこなう方法もあろう。

（１３）文書作成支援、検索、データベース

音声ワープロの目的は文書入力ではなく文書作成であるから、文書作成を支援する編集機能、辞書等のデータベース検索機能、さらにはユーザのアイデアを発展させる発想支援機能が求められる。

（１４）ユーザインタフェース

単語認識率98％以上の認識性能が望まれるが、誤認識は避けられない。したがって、候補からの選択や誤認識結果の修正を簡単におこなえるユーザインタフェースが極めて重要である。また文書作成では入力作業とともに入力した文書の編集作業が必須である。このためにキーボードやペン入力、さらに音声出力を同時に使用した使いやすいマルチモーダルインタフェースが望まれる。特に携帯情報機器はキーボードレス化、ペン入力の傾向にあり、音声による文書入力とペンによる編集というそれぞれの長所を併せたユーザインタフェースが有望である。また、多量の文書のすべてを音声入力することはユーザにとって負担が大きいため、音声入力とキーボード入力の両方の入力手段を用意し、ユーザが選択できるようにしておくことも重要である。

キーボードの苦手な人や身体障害者が使用する場合には、音声コマンドを用いた編集のためのユーザインタフェースが必要である。この場合、文書入力と音声コマンドとを区別するため

に音声コマンドには特別のキーワードを一緒に入力したり、音声コマンドによって入力モードと編集モードとを切換えることが必要である。身体障害者が使用する場合や電話口述筆記の場合にはさらに、音声出力による入力文書の読み上げ、操作確認、操作ガイダンスの機能が必要になる。

また、文書の編集内容をユーザが指示するだけでなく、音声ワープロが編集の仕方についてのガイダンスをおこなったり、ユーザと音声ワープロの協調作業で文書作成をおこなう対話型インタフェースが望まれる。

（１５）ユーザインタフェース評価

音声ワープロでは文書入力と同時に文書編集をおこない、さらにユーザインタフェース音声、キーボード、ペン入力などを組み合わせたマルチモーダルインタフェースになるため、使いやすさは単なる音声認識性能だけで決めることはできない。文書作成の効率、ユーザの負担の程度を総合的に評価する方法の確立が必要である。

3.1.4　音声ワープロの研究例

(a)NTT[山田 92]

かな・漢字のtrigramを利用した日本語dictationシステムの開発を行なっている。漢字、かなといった文字の連鎖確率情報を用いることによって、

(1) かな文字のみに基づく連鎖モデルに比べ、長い音素系列の統計量を扱うことができる、
(2) 後処理にかな漢字変換が不要となる、
(3) 文字の種類が増大しても、かな・漢字連鎖モデルの方がかな文字あたりのパープレキシティが小さくなる、

という利点がある。

国際会議の問い合わせに関するテキストデータベースを用いた評価実験を行なっている。約10万文節のテキストでかな・漢字trigramを学習し、274文節の評価用テキストの入力を行なった。音声認識には男性話者1名で学習した音韻HMMと予測LRパーザを用いている。特定話者システムである。実験の結果、文節変換率65.0%、文字変換率78.5%という結果が得られている。この他に医療の分野における医用所見作成をタスクとした実験も行なわれている[松永 92]。

(b)DARPA

音声ワープロの研究例としてDARPAプロジェクトにおけるWall Street Journal(WSJ)の記事を対象にした連続音声認識システム研究について紹介する。

DARPAプロジェクトではWSJを認識タスクとして大語彙連続音声認識を目指した方式の研究を開始している。この研究では語彙5000語および20000語の認識システムを目指している。Dragon、CMU、Lincoln Lab.、SRIで研究が進められている。いずれもDARPAプロジェクトのリソースマネージメントタスクあるいはATISタスク向けに構築されたシステムを大語彙連続向けに改良している。大語彙化に伴う計算量増加を如何に避けながら認識性能を確保するかが、現在の主な研究課題である。

Dragon[Baker 92]ではTied-Mixture HMMがパラメータのモデル化に効果を上げ、認識性能を向上させている。特定話者の5000語タスクに対して、従来モデルでは単語エラー率が35.6%であったのが、tied-mixtureモデルによって単語エラー率14.8%に改善されている。

CMU[Alleva 92]はSPHINX-IIをWSJタスクに適用した。SPHINX-IIでは単語間調音結合モデル、高次ダイナミックス、男女別の疑似連続HMM、共有分散モデルを導入することによってはSPHINXより単語エラーを50%以上低減している。WSJタスクに適用するに当たり計算量削減のために、単語モデルの文脈依存を左側だけに限定し、またViterbiサーチを改良した。

これによって処理時間が実時間の50倍以下となった。不特定話者の5000語タスクに対して18.4%、20000語タスクに対して32.7%の単語エラー率を得ている。

Lincoln Lab.[Paul 92]はTied-Mixture HMMを改良した。時間同期ビームサーチは音響高速マッチ(acoustic fast match)を用いたスタックデコーダ・サーチに変更された。5000語タスクで不特定話者15.0%、20000語タスクで特定話者13.6%の単語エラー率を得ている。

SRI[Murveit 92]は、ATISタスクに対して用いられたものと本質的に同一のDECIPHERシステムをWSJタスクに適用した。不特定話者5000語タスクに対して17.1%の単語エラー率を得ている。

3.1.5 まとめ

音声ワープロの使用形態と特性、実現のために必要となる技術要素と研究課題について考察した。音声ワープロに必要な技術要素のほとんどにおいて現状の技術レベルとのギャップが大きく、その実現は今後の研究成果によるところが大きい。

最後に、上記で述べた技術要素、研究課題の主なものに対して5年後、10年後に予想される技術水準を表3.1-2に示す。

表3.1-2　音声ワープロの技術要素と研究課題

技術要素、研究課題	5年後	10年後
音素認識	ていねいな発声が必要	普通の発声
大語彙単語認識	20,000語	200,000語
連続音声認識	文節発声	任意単位発声
未知語入力	未知語登録	未知語自動検出
音声対話	単語ベース	自由発声
話者モデル学習・適応化	教師あり事前学習	教師なしオンライン
言語モデル学習・適応化	事前学習	オンライン学習
耐雑音認識	定常雑音対応	非定常雑音対応
電話音声認識	限定語彙、文法	大語彙、自由発声
低コスト音声認識	ボード+ソフト	チップ、携帯機
遠隔マイク入力	近接マイク	ハンズフリーマイク
話者識別	テキスト依存	テキスト独立
文書作成支援、検索、データベース	文書編集	発想支援
ユーザインタフェース	ユーザ指示型	対話型

参考文献

1.[山田 92] 山田, 松永, 鹿野: "読みの情報を付与したかな・漢字文字連鎖情報による統計的言語モデル", 日本音響学会講演論文集, pp.121-122, 1992.3

2.[松永 92] 松永, 坪井, 山田, 鹿野, 佐藤: "かな・漢字統計量に基づく医用所見の認識", 日本音響学会講演論文集, pp.103-104, 1992.3

3.[Baker 92] J. Baker et al., "Large Vocabulary Recognition of Wall Street Journal Sentences at Dragon Systems", Proc. of the DARPA Speech and Natural

Language Workshop, pp.387-392, 1992.2

4.[Alleva 92] F. Alleva et al., "Applying SPHINX-II to the DARPA Wall Street Journal CSR Task", Proc. of the DARPA Speech and Natural Language Workshop, pp.393-398, 1992.2

5.[Paul 92] D. B. Paul, "The Lincoln Large-Vocabulary HMM CSR", Proc. of the DARPA Speech and Natural Language Workshop, pp.399-404, 1992.2

6.[Murveit 92] H. Murveit et al., "Performance of SRI's DECIPHER Speech Recognition System on DARPA's CSR Task", Proc. of the DARPA Speech and Natural Language Workshop, pp.410-414, 1992.2

3.2 福祉ロボット

3.2.1 背 景

現在、産業用ロボットは生産性の向上になくてはならない存在であが、将来は一般社会においても、身近なところでロボットが使われるようになると予想される。特に、医療・福祉の分野へのロボットの導入が期待されている。現在、高齢化社会が急速に進行しつつあり、介護者の不足が心配されている。寝たきり老人のほとんどは家庭の中で生活をしており、家族に大きな負担がかかっている。そのため、老人や障害者を看護したり、身の回りの世話をする福祉用ロボットの開発は重要である。また、ロボットの利用は介助者の負担を軽減するだけでなく、介護を受ける側も気兼ねなく世話を受けられる利点がある。また、将来、家事ロボットが技術的に可能になれば、老人世帯の家庭で、食事の用意、掃除、洗濯などのルーチンワークのロボット化が望まれるようになると思われる。

このようなロボットに指示を与えたり、協調して作業をおこなううえで、音声によるインタフェースは体の不自由な人や機械操作の苦手な高齢者に適している。以下では、このような福祉ロボットの音声による制御に関し、必要とされる課題について考察する。

3.2.2 利用形態

福祉ロボットの形態とその開発例を述べる[中野 82]。

（1）日常生活介助ロボット

これは寝たきり老人や障害者の日常生活を介助するロボットである。具体的には、物品の運搬、食事の世話、電話の利用、電動ベッドや空調の調節、カーテンの開閉、テレビの選局、電灯のオンオフ等が挙げられる。これらの中、各種機器の操作は、現在でも音声リモコンで実現可能であが、マニピュレータ等の動きを必要とする機能は複雑で、マンマシン対話方式によるインタフェースが必要となる。

電動ベッドを音声で制御する試みはすでに商用システムとして試用段階にあるものもある。米国の病院で試用した例では［Liang 89］、特定話者の孤立単語音声認識により音声でベッドの状態を制御している。この例では、食事、読書、ＴＶ等の動作目的によりマクロ命令であらかじめルーチン化された動作を選択し、さらにミクロな命令で制御できるようになっている。命令の語彙数は10語以下で、他人が操作できないように、パスワードが使用されている。

（2）障害者搬送ロボット

身障者や患者をベッドから抱き上げて移送したり、入浴させたりする世話を介助ロボットで代替できれば、介護者の負担を軽減することができる。このようなロボットでは、通常は介護

人が操縦し、夜間等の介護人が不在の場合に患者自身が簡単な指令で操縦できるように、素人にも容易な対話方式によるマンマシンインタフェースが不可欠である。このようなロボットとして、全方向の移動機構、人間抱き上げマニピュレータ等の機構部分については試作された例も報告されている［中野 1981］。

筋ジストロフィーや脳血管障害のある人では、車椅子の運転も困難なため、病院や家庭内での移動にも手助けが必要となるので、音声で車椅子が運転できれば介護者のみならず患者自身にとっても極めて有益である。最近、このような音声制御機能のついた商用の車椅子が発表されている。この車椅子は、「進め」、「止まれ」、「後ろ」、「右」、「左」の5単語により制御するもので、音声認識システムとしては非常に簡単なものである。

この他、障害者を車椅子や浴槽に移送する実用的な手段として、すでに各種のリフターが商品化されているが、これを音声で制御できれば介助者の手助けが不要になる。

（3）盲導犬ロボット

盲導犬に適した犬はごくわずかで、1匹の犬を盲導犬に仕立てるのには厳しい訓練と莫大な費用がいる。また個人が盲導犬を保有する場合、その維持と世話に非常な手間と経費がかかる。そのため、盲導犬の機能をロボットで代替しようというのが盲導犬ロボットである[館 78]。このようなロボットには、盲人の命令に従って、直進、左折、右折などの行動を先導したり、障害物がある場合に命令に逆らって停止する機能が要求される。このとき、ロボットと人間とのマンマシンコミニュケーションに音声が併用できれば、両手が自由になり、またより多くの命令を使えるようになると期待される。

以上のような利用形態のうち、移動が可能で複数の複雑な作業をするロボットでは、かなり高度の知的制御が要求される。特に、高齢者や障害者にも容易に使えるためには、ヒューマンフレンドリーなロボットであることが必要であり、そのためには、マクロな命令でその時のタスクと状況を判断しながら自立して行動し、必要に応じて利用者と対話による確認をしながら動作することが要求される。また、ロボットの誤動作で利用者に危害を加えることのないよう、フェイルセイフなシステムであることも重要である。

3.2.3 音声認識部の機能・性能

（1）話　者

通常、福祉ロボットの利用者は特定の話者に限られるが、複数の利用者が共同で使用したり、介護者と患者の双方がロボットを操作する場合も考えられる。また、利用者が一人に限られる場合であっても、高齢者や病人では音声を登録するだけでも困難が予想される。したがって、

基本的には不特定話者のシステムが望ましく、さらに、少数の音声による話者適応かオンライン適応機能を備えることが必要であろう。

一方、複数話者の音声が存在するような環境でロボットを使用する場合には、特定の利用者の音声のみに選択的に応答するような、話者認識機能が要求される。このときは、当然テキスト独立型の話者認識が必要となる。

（2）語　彙

ロボットの制御に使われる語彙は少ない。ロボットの中でも家庭における日常生活を介助するロボットの場合は、作業内容が多いので全体としての語彙数は比較的多い方であるが、この場合も、作業内容によりマクロコマンドで語彙を階層化して制限できるので、実質上の語彙は少数で間に合うと考えられる。しかし、少ない発声で滑らかな制御をするためには、単に「右」「左」等のディジタル的な動きを制御する単語ではなく、「もう少し右」等の程度表現する単語も認識する必要がある。

（3）発話様式

ヒューマンフレンドリーな制御をするためには、孤立単語ではなく自然発話を許容する必要がある。特に、機械に不慣れな高齢者や素人の介護者では、不要語、言い直し、言い淀みの頻度が多いと予想される。

（4）情緒性情報の認識

音声に含まれる情報を最大限に利用し、迅速かつスムーズな動作をさせるためには、言語情報だけでなく情緒的情報も抽出し、患者の感情や心理状態をある程度理解できることが必要である。たとえば、緊急時の声や苦痛を訴える声などである。

（5）対話機能

老人や障害者にとって使い易いシステムであるためには、ロボットに与えるコマンドは少数のマクロなコマンドとし、介護者が患者を看護するときのように、ロボット主導で対話により確認をとりながら作業を実行することが必要とされる。すなわち、作業内容（タスク）を指示するまでは利用者主導で、それ以後はロボット主導の対話を行い、また、場合によっては途中で利用者に主導権を戻すことができるようなシステム構成が望ましい。また、人の介護を対象としたロボットでは、認識に即応性とフェイルセイフも要求されるので、ロボットからの問い合わせの途中であっても、音声による割り込みで動作の中断や変更のできることが必要である。

（6）入力方式

家庭や病院におけるロボット利用では、接話マイクは不適当で、ベッド等に固定した遠隔マイクが最も好ましい。また、それが困難な場合であっても、少なくとも衣類等に固定したワイヤレスマイクが必要である。この他、電話を介して家の外から指示を与える場合も考えられる。

（7）耐環境性

介助ロボットの場合には、通常、家庭や病院等の比較的静寂な環境で使用されるが、利用者が老人や病人で音声レベル自体が低い場合には、やはりＳＮＲの低下が問題になると予想される。またこのような環境では、エアコンや掃除機の定常的な騒音だけでなく、テレビ、ラジオ、人の声等のスペクトル変動の大きな騒音に強いシステムである必要がある。また、車椅子や、盲導犬ロボットのように屋外で使用するロボットの場合には、交通騒音のような比較的大きな変動騒音下においても、十分な認識精度を維持する必要がある。

また、ロボットの制御に遠隔マイクを使う場合、マイクロホンと口との距離・方向が変化し、音響伝送歪みやレベル変動が問題となる。したがって、このような変動に対する耐性を高め、スムーズな正しい動作ができるだけの認識精度を維持しなければならない。

3.2.4　研究課題

以上のような、音声によるロボットの制御で要求される研究課題として、次の項目が挙げられる。

（1）高齢者音声の収集

高齢者の音声や弱々しい声、介護・看護における対話音声等のデータベースの収集が必要である。現在、老人や病弱な人の音声のスペクトル特性やその変動特性についてほとんど調べられていない。また、実際の看護や介護で使用される言葉を収集し、「右」「左」等の限定された語彙ではなく、日常使われる言葉をできるだけ多く取り入れる必要がある。

（2）ロボット用対話処理

ロボットとの対話は制御に関する確認が主で、比較的単純である。しかし、情報案内等における対話と異なる点として、発話の理解に、文脈や作業内容（タスク）に関する知識だけでなく、ロボットの動作・移動に伴う外界の状況に関する情報が利用される点が挙げられる。すなわち、図3.2-1に示すように、位置、圧力、画像等の各種センサーからのマルチモーダルな情報を用いた対話処理の研究をおこなう必要がある。

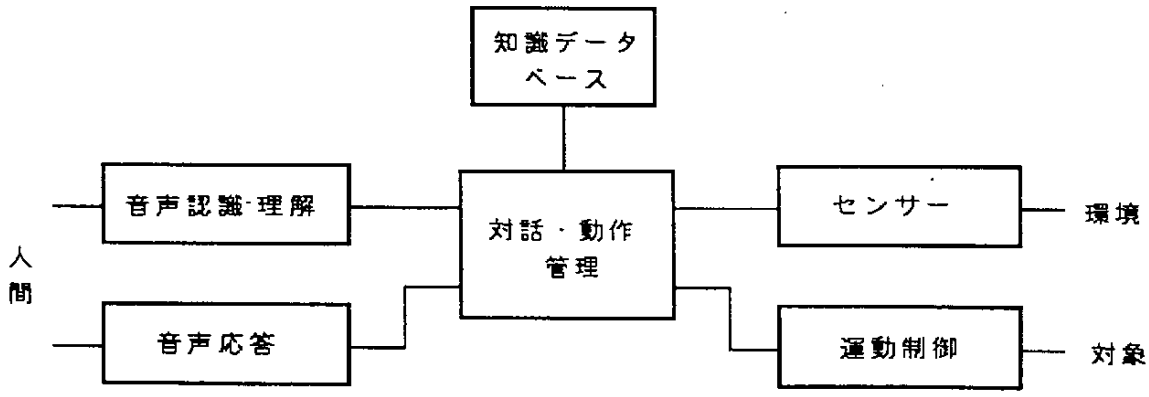


図3.2-1　音声ロボットの制御

（3）自然発話の認識

ロボットの制御に関する対話では、どの様なコマンドが発声されたかがわかればよく、また語彙もあまり多くないので、自然発話で問題となる不要語や言い淀みに対処するのに、ワードスポッティング手法が有効と考えられる。このような少数語彙の対話システムの良い例として自動注文システムがあげられる[Takebayashi 93]。ワードスポッティングでは短い単語の部分マッチングや耐雑音性の問題について更に研究を進める必要がある。

（4）雑音・伝送歪みへの対策

現在、定常騒音や特定の変動騒音に対しては、音声と雑音のHMMの分解・合成による耐雑音認識手法が有力視されているが[Varga 90][Gale 92]、人の声や音楽のようにスペクトル変動の複雑な騒音に対してはまだ解決すべき問題が多い。また、レベルや音響伝送特性の影響に関し基礎的な実験例はあるが[飛田 90]、そのような変動に対して頑健な認識手法はまだ確立されていない。

（5）話者適応

ロボットの制御に使われる語彙は比較的少ないので、不特定話者の認識は容易と思われる。しかし、高齢者の声を含む多様な音声を対象とするので、マルチテンプレートか話者適応の手法が必要になると考えられる。特に、話者適応においては、利用者に負担をかけないシステムとするため、少数の音声による教師無し話者適応やオンライン適応手法を開発する必要がある。

（6）話者認識

複数の音声から特定話者の音声区分を自動抽出する方法を開発する必要がある。これは、テキスト独立型の話者認識の問題で、現在、ベクトル量子化[Soong 87]や混合ガウス分布を用いたテキスト独立な認識手法[Tseng 92]が開発され、認識精度は向上している。しかし、まだ高精度の認識には数秒以上の音声を必要としている。したがって、今後より短時間の音声で認識可能な手法と時間的連続性を考慮した同一話者区間の検出法を開発する必要がある。

（7）情緒性情報の処理

発話に含まれる言語以外の情報から、命令の程度（強さ）、緊急性、満足度等の判断を可能とするため、情緒性と音声の物理的特徴との関連を明らかにする必要がある。たとえば、同じコマンドでも発声速度、レベル、ピッチの違いにより要求の強さや緊急性などもある程度判断できると予想される。

3.2.5 むすび

以上のように、福祉ロボットにおける音声認識のタスクは命令型の比較的単純なものであるが、命令を実行すべき動作の対象が人間であるために安全性と正確さが要求される。また利用者が老人や障害者であるためヒューマンフレンドリーなインタフェースが要求される。そのため、ロボット制御の音声認識では、自然発話、耐雑音性など実用上の課題が多い。最後に、これらの課題に関する今後の技術水準の予測を表3.2-1に示す。

表3.2-1　音声によるロボット制御の技術予測

	現　在	5年後	10年後
対話形式	リモコン	利用者主導 簡単な対話	ロボット主導 簡単な対話
発話形態	孤立単語	連続単語	自然発話
話　者	特定（登録）	グループ内 不特定	オンライン 話者適応
話者限定	手動切り替え	テキスト依存型	テキスト独立型
耐雑音環境	定常騒音 中SNR	定常騒音 低SNR	特定の変動騒音
遠隔マイク	不　可	静寂な部屋で可	可

真に使い易い、ロボットを実現するためには、ロボット自身に関し、他のセンサー情報を含めた高度の知的制御の実現を待たなければならないが、音声インタフェースが不可欠な場合も多く、今後音声入力の役割が期待される。

参考文献

1.[Gale 92] M.Gale and S.Young, An Improved Approach to the Hidden Markov Model decomposition of Speech and Noise, Proc. ICASSP92, I, pp.233-236, 1992.

2.[Liang 89] M.D.Liang and K.Narayanam, "The Application of Voice Recognition to Robotic Positioning of a Hospital Bed," Speech Technology, pp.30-33, 1989.

3.[Nakano 81] E. Nakano, et al., First approach to the development of the patiet care robot, Proc. 11th Inter. Symp. Industrial Robots, pp.87-94, 1981.

4.[中野 82] 中野栄二、「福祉ロボット」、電子通信学会誌 Vol.65, 4, pp.443-449 (1982)

5.[佐藤 92] 佐藤倫太、開一夫、安西祐一郎、「ロボットとの対話：センサ情報を利用した音声対話システムLintaの設計と実装」、情報処理学会研究グループ資料、92SLP-1-3, pp.19-26 (1992).

6.[Soong 87] F.K.Soong, A.E.Rosenberg,B.H.Juang, and L.R.Rabiner, A vector quantization approach to speaker recognition, AT & T Tech. Journal, 66(2), pp.14-26, 1987.

7.[Takebayashi 93] Y.Takbayashi, H.Tsuboi, H.Kanazwa, Y.Sadamoto, H.Hashimoto, and H.Shinchi, A real-time speech dialogue system using spontneos speech understandng, IEICE Trans. INF. & SYST., vol.E76-D, No.1, pp.112-120, 1993.

8.[館 78] 館、他、人と一定間隔を保つ移動ロボット制御の一方法、バイオメカニズム、4、pp.279-102 (1978-06)

9.[飛田 90] 飛田瑞広，菅村 昇，中津良平，音響伝送歪みの単語認識率への影響とその改善法,信学論(D), J73-D, 6, pp.781-787, 1990.

10.[Tseng 92] B.L.Tseng, F.K.Soong, and A.E.Rosenberg, Continuous probabilistic acoustic Map for speaker recognition, Proc. ICASSP92, II, pp.161-164, 1992.

3.3 CAIシステム

3.3.1 背 景

CAI はすでに30年以上の歴史を有している。初期の代表的なシステムとしては米国イリノイ大学のPLATO（Programmed Logic for Automatic Teaching Operation）があり、1960年に1台の端末をもった研究システムとしてスタートしている。1970年代に入ると計算機の発達もあり、大型計算機によるCAIシステムが全盛となり、1980年代のパーソナルコンピュータの急速な進歩によって、教育現場に急速に浸透してきている。基本的な機能はドリル型、ゲーム型、シミュレーション型、説明型等に分類され、教材、教授方法自体は初期に開発されたものと大きな相違はないものの、LAN型CAIに、楽器やビデオなどの教材が接続可能となったこと、大規模な学習用データベースが構築されるようになったこと等により、教材の提示の観点では、イメージ、動画、音声の利用など方式も多様になり、大きく進歩してきている。しかし、このような従来のCAIシステム（伝統的CAI）は、使い込まれるうちに多くの欠点も指摘されるようになっている。主要な問題点としては、

(1) 学習者が疑問点を質問できない、

(2) 入力方法が穴埋めや、選択、数式、単語などの特殊な方法に限られる、

(3) 出題された問題に対しての正答、誤答、ヒントをあらかじめ用意する必要がある、

(4) なぜ誤ったかをシステムが推定できない、

などが挙げられ、CAIが「学習者を受け身にする高価な自動ページめくり機」との批判を受ける結果となっている。CAIの研究ではこれらの問題点を解決することが大きな課題であり、現在、知識処理にもとづき学習者に応じた指導をおこなう事を目的とする、いわゆる「知的CAI」の研究が進められている。上記の問題点の(1)(2)の観点から自然言語インタフェースは知的CAIには必須の問題と考えられ、音声認識·理解技術、音声合成技術への期待も大きい。

また、音声そのものが学習の対象である「言語学習システム」は音声認識·理解、合成技術により初めて実現可能で、今後のグローバリゼーションに伴い言語教育の必要性が増す中で、言語学習CAIシステムへの要求もますます高まってくるものと考えられる。

さらに、従来型のCAIシステムのインタフェースに音声を利用する効果も高く、シミュレータ型CAIの発展として、音声対話あるいは発話による指示を伴う業務の訓練システムも音声認識·理解技術により訓練の高度化、効率化が可能となる。

3.3.2 音声インタフェースを有するＣＡＩシステムの形態

（1）従来型CAIの音声インタフェース

従来型のCAIでは、学習者は質問に対しての回答を、単語や数字、「はい」、「いいえ」等の語で入力する。したがって、音声認識を用いる場合は、基本的に離散単語音声認識となる。

学習者はCAIシステムに対して、音声、キーボード、マウス、ペンなどを用いて入力し学習を進める形態となる。音声認識は操作が容易であること、思考の中断が少ないことから効果的である。さらには、音声認識システム自体の目新しさやゲーム的要素は、他のマルチメディア機能（オーディオ出力、音声メール、映像表示、コンピュータグラフィックスなど）と相まって学習者の興味を引き、やる気や興味の持続力の向上にも役立つと考えられる。

このような従来型システムでは、CAIシステム自体が学習をシステム主導で進めることから、音声認識によるインタフェースは比較的容易に実現が可能である。しかしながら、効果的なシステムを構築するためには、次のような高機能な音声認識が必要である。まず、CAIシステムでは学習者が思考中に発話するため、付帯語・言い直し等を含む曖昧な発声が多くなることが考えられ、これらに対処できる音声認識機能が求められる。

さらに、設問形式が選択回答法の場合には、少数限定語彙の単語認識が利用できるが、自由語による回答方法では未知語処理や意味カテゴリの判別・同定などの高度な機能を備えた大語彙認識が要求される。

（2）言語学習システム

言語学習をおこなうCAIシステムでは学習対象が音声言語そのものであることから、音声認識合成技術は必須である。たとえば発声訓練システムでは、システムが提示した語あるいは文章を、学習者が正しく発声しているか否かを評価検定する機能が要求され、音韻性（読みが正しいか）、韻律性（語のアクセント、文イントネーション、リズムが正しく自然か）の両面が扱われる。また効率的な学習のためには、単に模範音声を提示したり口の動きを見せるだけでなく、音声の特徴量（ピッチパターンなど）のビジュアライゼーションも有効であり、さらに学習者の発声上の問題点を的確に指摘する必要もある。

次に言語表現の訓練では、学習者の自発的な発話を扱えることが望ましく、前記のような音響的側面にとどまらず、構文・文脈の評価機能も必要となる。状況意味論などの言語処理を含め音声理解技術を高度利用した知的CAIシステムとして将来が期待される。

（3）音声を用いる業務の訓練システム

音声による命令指示や音声対話を必然的に伴う業務では、シミュレーション型のCAIシステムに対しての要求も強い。これらの例としては、航空管制業務、複数のクルーで操縦する航空機の操縦、船舶の操船、プラントの運転・監視、各種保守・補修等が挙げられる。様々な状況のもとでの訓練をおこなうためにはシミュレーション型のCAIシステムは極めて有効ではあるが、複数の訓練者あるいは教官が揃わなければ訓練ができない、さらには人間が対応するため均質な訓練ができない等の問題がある。音声認識の導入によりこれらの問題点を解決できるほ

かに、訓練者の取った行動（発声）の詳細なログが残るため、従来人間が対応していた場合に比べ、訓練内容の事後照査、詳細な指導が可能になるものと考えられる。

これらの実現には多くの場合、連続音声認識が要求される。そのタスクについて考えてみると、音声による指示を含む業務ではヒューマンエラーを低減させるため、多くの場合発声についての比較的詳細な取り決めがある。したがって要求される連続音声認識の対象構文は人工構文に近いものであり、また、文の意味も曖昧性が少なく、言語処理の問題からは比較的容易なタスクであると考えられる。ただしシミュレータの現実感を高めるためには環境騒音等も現実のものに近い形にすることが要求されること、訓練員自体が複数である場合など、騒音耐性が要求され、認識率自体も極めて高いものが要求される。

（4）知的CAIシステム

知的CAIシステムは前述の従来CAIの問題点を解決するものとして研究されているシステムであり、個別指導等のCAIシステムとしての機能もその研究対象であるが、音声認識利用の立場からは、

・自然言語インタフェースを有する

・学習者が質問できる（双方向主導対話システム）

システムとして考えられる。さらには音声を導入することで、学習者の発声の言い淀みや、言い直し、発声速度などの韻律的な情報から、学習者の理解の程度をより詳細に推定する可能性が生まれる。

上記(1)(2)項に、知的CAIにおいて必要とされる音声認識機能の例を述べている。

3.3.3 技術課題

上記の各システムを実現するための技術課題を以下に示す。ただし、基本となる連続音声認識技術等については簡単に述べ、CAIへの適用を考えたときの特異な問題を中心に示す。

（1）話者適応化技術

CAIシステムは他の音声認識システムに比べても、利用者が比較的長時間使用する形態をとる。また、学習者が混乱を起こさないためには、音声認識が極めて高い認識率で動作することが求められる。このような観点から、話者適応化技術はCAIシステムの音声認識には必須であると考えられる。利用形態によっては、教師付き学習をおこなわせることも可能ではあるが、教師無しであることが望ましく、言語学習システムへの適用の場合などでは、訓練用テキストの発声そのものが不確実な学習者も想定され、教師無しであることが必須となる。

（2）付帯語や無意味語を含む離散単語音声認識

学習者が思考し回答を発声するという過程では、言い淀み、言い直し、無意味語発声が生じる。システム利用時にある程度の発声方法についての注意事項を利用者に示すこともできるが、完全に避けられる問題ではなく、また、発声法についての強制をおこなうことは、思考の中断がないという音声認識のもつ利点が活かされないことになり、利用者にストレスを与えるのみとなる。この分野では最近、無意味発声のHMMを単語モデルに接続した方式など、音声認識の実用化課題として研究がおこなわれている。今後は、さらに高精度化していくためには、これらの発声を含むデータベースの収集も不可欠である。

（3）語の意味同定

発声された語がシステムに指定された単語カテゴリーと意味的に同一であるか否か、あるいはどの程度近いかを判別する機能は、知的CAIでは必須技術であり、また従来型CAI、言語学習システムにおいても効果的システム構築のために有効である。離散単語音声認識の未知語のリジェクト問題としての検討はおこなわれているものの、意味論的な検討は今後の課題である。

（4）韻律処理（アクセントの識別など）

日本語の言語学習システムでは音韻性（読み、正しい発音の学習）の他に、韻律性（正しいアクセント、きれいなリズムの学習）の処理が必要となる。単語のアクセントを誤っていないかを識別する技術や、ビジュアライゼーション、また誤りや不適切な発声に対しての適切な指示の提示が課題となる。この課題も、一部の聴覚サイドからのアプローチや初歩的な言語学習システムでの検討などがおこなわれているのみで、言語学習CAIシステムの立場からの本格的な検討がなされていない分野である。

（5）連続音声認識

音声対話型の業務の訓練システムでは多くの場合、すでに文形等についての比較的詳細な制約事項が決まっている場合が多い。これらに対しての連続音声認識の言語処理部は、比較的簡単なモデルである有限オートマトンや単語nグラム等の方式で対処できる範疇であると考えられる。

（6）自然会話文章の認識

学習者は、システムの質問に対して、思考し、答え、また質問をおこなう。このような過程での発声は多くの場合、言い淀みや、言い直し、無意味語を含み、構文的にもill-formedとなる。これらに対しての言語処理技術の開発が必要である。部分パーシング等の手法や、意味

スロットフィリング等いくつかの試みがおこなわれているが、知的CAIに望まれている学習者の質問という、意味構造を規定しにくい自由発話の会話理解は極めて困難な問題である。

（7）言語情報以外の利用

通常の目的達成型の音声認識では入力される音声の言語情報のみが意味をもつ。しかし、学習時の教師と学習者の間の対話を考えると、質問から回答までの時間、発話中の言い直しや言い淀み、さらには「自信の無い言い方」といったある種の情緒性が学習者の理解の程度を推定する重要な情報になり得る。情緒性の判定などは個人性の問題もあり極めて困難な問題ではあるが、音声利用の大きな利点でもあり、利用できる情報の学習者モデルへの反映は検討されるべきである。

（8）対話モデル

従来型のCAIシステムにおいては対話はシステム主導でおこなわれる。したがって複雑な対話モデルを必要とはしない。しかし、対話の枠組み自体が学習の進め方そのものであることから、システムを構築するにあたっては、学習手順の知識源がそのまま音声対話のデータベースとして利用できる必要がある。

プラントの運転や、航空管制などのシミュレーション型の訓練システムの場合は、対話は比較的明確な目的達成のためにおこなわれるものであり、やはりそれほど複雑ではないと考えられるが、プラントの状況や、航空機の位置などの動的な知識源に依存する。このようなタスクに依存した動的な知識源の有効な利用方式が検討されるべきである。これらの研究では人間同士が業務を遂行しているときの対話データベースの構築も重要である。

知的CAIの目標とする双方向主導の対話を実現することは極めて困難である。システム主導での動作については、効果的な学習が可能な指導方略と、その時点での学習者の理解の程度を適切に表す学習者モデル、学習者の誤りの原因同定機能、学習者の特徴と理解知識から適切な学習対象や指導方略を選択あるいは変更する機能といった、CAIシステム自体の課題が解決されねばならない。その上で音声理解システムが教材知識や教授知識を知識源として動作できるようにする必要がある。学習者主導の対話ではさらに、意味構造の規定が困難な対象文章を、教材知識を知識源として理解する高度な音声理解が求められることになる。

（9）知識源、システム構築

知的CAIを例とすれば基本的な知識源としての教材知識、教授知識、さらに動的な知識源である学習者モデル、これらの表現法と音声認識の知識源としての利用方式の検討が必要となる。知識源の矛盾が生じないようチェックする機能等も必要となる。

（１０）言語生成とテキスト音声変換

教材知識、学習者モデルにもとづく解説文、質問文を出力するため、概念からの文生成技術が必要となる。また、テキスト音声変換方式については、自然で高品質であることはもちろん、教示の意図に応じての強調や感情表現が可能であることが望ましく、意味情報からの多様な表現が可能な合成方式の研究が課題となる。

（１１）システムの評価

CAIシステムの評価は極めて困難な問題であり学習効率の尺度も多種多様な評価基準が考えられる。そもそもCAIシステム自体は決して人間教師の代わりにはならないものであり、人間教師のように、教育技術だけでなく優しさ、尊敬、熱意等を含めた「良い先生」を目指すものではない。したがって、なぜCAIを導入するかという目的に応じた評価尺度があるべきであり、音声認識を導入した効果もその尺度での評価をおこなう必要がある。

３.３.４　まとめ

CAIシステムについての知的音声処理の形態と技術課題について示した。最終的な音声対話知的CAIシステムの達成に向けてはCAIシステム自体の課題も含めて技術課題は山積している。５年ないし10年後をイメージすると、上記課題の部分的な解決により以下のようなシステムの構築が可能であると考えられる。

［５年後］

◆従来型CAIシステム

・発声：離散単語（付帯語、無意味語等の処理を含む）

・語彙：少数語彙（正解の選択）、語の同定、大語彙（データベースの検索）

・話者適応化機能

・対話：システム主導、単純なモデル

・システム全体の形式：マルチモーダル、マルチメディア、LAN分散環境、ドリル型、プラクティス型

◆シミュレーション型CAI

・発声：自由度の低い（業務の基準にもとづく）連続音声

・語彙：3000単語程度

・話者適応化機能

・対話：単純なモデル、動的知識源の利用、簡単な学習者モデル

・システム全体の形式：実時間シミュレータ、バーチャルリアリティの利用等

◆言語学習（システム）

・発声：離散単語、単純な文

・認識：単語認識。語、文の同定

・韻律処理：単語アクセントの同定

・システムの機能：単語の発音アクセント指導。音声特徴のビジュアライゼーション

［10年後］

◆言語学習（システム）

・発声：単語、文

・認識：単語認識、限定した文を対象として連続音声認識。
語、文の同定。学習者からの簡単な質問を許す。

・韻律処理：単語アクセント、文イントネーションの同定、評価。

・機能：発音の指導、韻律の指導。システムと対話することでの会話訓練。

◆知的CAI

・発声：連続音声（協力的な発声などシステム利用にあたってのある程度の発声の規約を事前に提示した上で起こる、非文法的文、言い淀み、言い直しの処理）

・語彙：数千語

・対話：基本的にはシステム主導（指導方略にもとづく制御）、部分的にユーザ主導を許す（学習者の質問について簡単のものを許容する）。

・言語情報以外の利用：単純な言い淀み等の情報を学習者モデルに適用

参考文献

1. ［大槻 88］大槻他："知的CAIのパラダイムと実現環境"、情報処理、Vol.29, No.11, pp.1255-1265, 1988
2. ［伊藤 88］伊藤："知的CAIシステム探訪"、情報処理、Vol.29, No.11, pp.1283-1293, 1988
3. ［竹内 88］竹内："知的CAIのフレームワークBOOKの構成といくつかの応用例"、情報処理、Vol.29, No.11, pp.1309-1315, 1988

3.4 情報案内システム

3.4.1 背 景

人は日常のコミュニケーションで音声を最も手軽でかつ容易なメディアとして用いている。人と機械とのコミュニケーションの場合はどうであろうか。人は、音声をもっとも身近なメディアとするが、機械はどうであろうか。ここでは、情報案内タスクを取り上げ、人（ユーザ）と機械（コンピュータ）のコミュニケーション（対話）について考察する。

情報案内タスクの特徴は、コンピュータが多量のデータベースを持っており、ユーザがそのデータベースを対話型で検索利用し、当初の目的を達成するという形態をとることである。通常、ユーザの検索によってデータベースが変更されることがないことも特徴と考えられる。典型的な情報案内タスクの例として、銀行の振込照会、窓口業務案内、地理案内、株価の問い合わせ、列車や飛行機の時刻表案内、電話番号案内、旅行案内などが考えられる。これらのタスクの問い合わせ事項は、比較的単純であり、コンピュータとの対話も単純な形態を取る。しかしながら、現在コンピュータが膨大なデータベースを持っており、コンピュータが日常生活で必要とされる多種多様なサービスを提供できるようになっている。このような状況であるので、情報案内システムのニーズは極めて大きくなってきている。

ここで、人による入力をモーダルと表現し、音声以外の入力手段を持つことをマルチモーダルと呼ぶことにする。また、コンピュータからの出力で、音声以外のメディアを持つことをマルチメディアと呼ぶことにする。マルチモーダル、マルチメディアの場合には、音声と他のメディアの競合が生じ、音声が相対的に使いよくなければならない。その意味で、音声認識、音声合成に対して厳しい性能が要求される。そのために、音声認識技術の飛躍的性能の向上、コンピュータの得意とするメディア（画像出力）の利用方法、認識誤りなどの不確実な情報に対処するための対話モデル、音声インタフェースの評価手法など、まだ、多くの研究を必要としている。

音声による情報検索の対話システムは、ANSER[Murakami 87][NTT 92]など語彙数が少ないものが、一部実用化されてきたが、HMMモデルと不特定話者音声認識技術の向上に伴い、株価の問い合わせ[Lennig 92b]などのように語彙数の多い分野でも用いられ始めている。これらは、マルチメディアではないけれども、電話を通したシステムとして使われ始めている。

ここでは、まず、音声による情報検索システムとして、数種類のタスクを取り上げ、その実現形態について考察する。特に、それらのシステムの実現に不可欠な基本要素について課題を明確にすることを試みる。実際のタスクとして、アメリカのDARPAプロジェクトのATISシステムを評価手法の例として取り上げる。さらに、NTTでおこなっている番号案内タスクを大語彙連続音声認識の例として取り上げて紹介する。最後に、音声による検索システムの研究課題についてのまとめを試みる。

3.4.2 音声による情報検索タスク

代表的な音声による情報検索タスクとして、（1）ANSERとして知られている銀行の振込や残高の照会、（2）自治体などでの窓口の業務の自動案内、（3）目的地まで道順等の地理案内、（4）株価の情報案内、（5）列車の時刻表や到着の案内サービス、（6）アメリカのDARPAプロジェクトで行なわれている航空機のフライト情報案内（ATIS）、（7）電話番号の案内[南 92][武田 92]、（8）旅行相談、などが考えられる。これらの情報案内サービスについて、(a) 認識対象に含まれる語彙数（小（～100）、中（100～1,000）、大（1,000～100,000））、(b) 発声の制約（孤立単語発声、連続単語発声、自由発声）、(c) 対話モデル（単純、普通、複雑）、(d) 入力の手段（音声認識のみ、マルチモーダル）、(e) 出力の手段（音声応答のみ、マルチメディア）について考えてみる。表3.4-1にその結果を示す。○はそのタスクで必要とされる技術や適する手段を示し、△は可能性のある技術や手段であることを示す。空白は、対応する技術や手段では、あまり可能性のないことを示す。

これらの結果を、横軸に音声認識技術の実現の容易性、縦軸にマルチモーダルとマルチメディアを取り、その実現領域を図3.4-1に示す。まず、メディアとして音声のみでも可能な（1）ANSER、（2）窓口業務案内、（4）株価案内と、マルチメディアとマルチモーダルが望まれるその他の（3）、（5）、（6）、（7）、（8）とに分れる。（1）、（2）、（4）のように既存の電話網を介した音声だけのメディアでもかなりのニーズがあることが推察できる[Voice 92][Billi 92][Harlow 92][Lennig 92a][Rosanbeck 92][Sorin 92][Wilpon 92]。

表3.4-1　情報案内タスクとその実現技術と手段

	語彙数			発声			対話モデル			入力(モーダル)		出力（メディア）	
	小	中	大	単語	連続	自由	単純	普通	複雑	音声のみ	マルチ	音声のみ	マルチ
ANSER	○			○	△		○			○	○	○	○
窓口業務案内	○	△		○		△	○	△		○	△	○	△
地理案内		○		△	△	○	○	△		○	○		○
株価問い合わせ		△	○	○	△	△	○			○	○	○	△
列車の案内	△	○	△	△	○	△	○	△		○	○	○	○
ATIS		○	△		△	○	△	○	△	○	△		○
電話番号案内		△	○	△	○	△	○	△		○	○	○	○
旅行相談		△	○		△	○		△	○	○	△	△	○

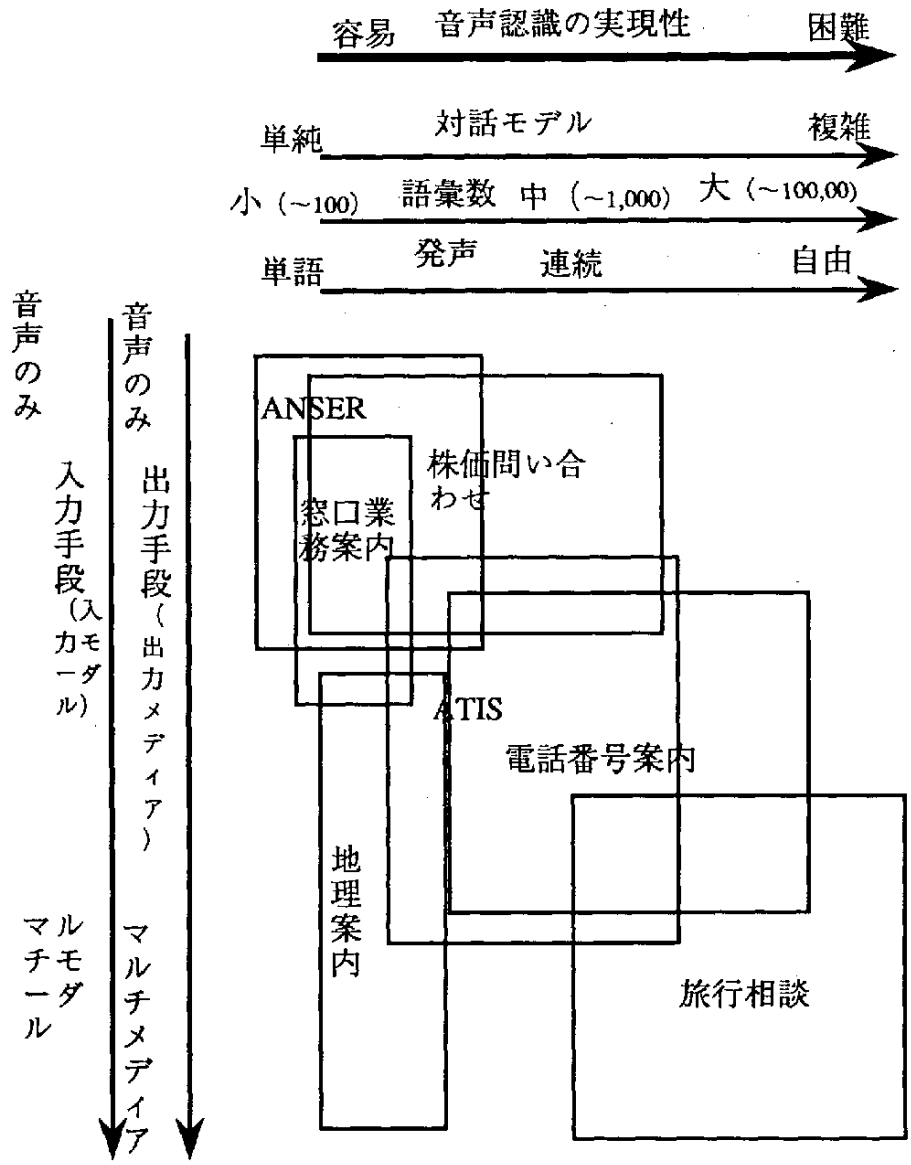


図３．４-１　情報案内タスクが必要とする技術

以下、音声認識の技術項目に重点をおいて、技術的課題について考察する。

（i）不特定話者音声認識

情報案内は比較的単純なタスクであり、コンピュータとのやりとりも比較的少ない。よって、ユーザがすぐに使えることが望ましい。しかしながら、あるユーザが様々のタスクに対して頻繁に音声入力を用いる時には、話者適応の技術も使われる可能性がある。電話のように音声だけの入力の場合には、すべての人でも動作させるという難しい問題を抱えている。実際には、すべての人で動作させるのは不可能であるので、人を使った代替のサービスを提供することが必要となる。しかしながら、マルチモーダルのパラダイムにおいては、ユーザ自身が入力手段

を選択することによって入力手段を換えることができるので、不特定話者の制約が若干ゆるくなる。

(ii) 雑音処理

通常の静かなオフィスで使われることが多いと考えられる。しかし、S／N比で10数dB程度の様々な雑音の存在には対処する必要がある。様々な雑音を含む音声データを集めて、雑音に対して頑健なHMM音韻モデルをあらかじめ作成することも、少数語彙の音声認識に対しては有効と考えられるが、認識精度を必要とする大語彙音声認識では、前処理としてサブストラクションなどの手法や、周囲雑音に合わせてHMMモデルを変更するHMM合成(HMM composition) [Martin 92]や、HMM分解(HMM decomposition)[Varga 90][Gale 92]の手法が有望である。また、コンピュータなどからの応答音声などをサプレスする、いわゆるバージイン(barge-in)の手法も、音響イコライザー[島田 92]の開発販売とともに可能となってきている。さらに、指向性マイクロフォンの利用や、複数のマイクロフォン[金田 92]を用意して、雑音をサプレスすることも有効である。その他、電話回線を通して音声を認識する場合には、電話のマイクロフォンの違いや、回線特性や回線雑音のばらつきに対処することも必要である。さらに、音声区間の抽出法や、ワードスポッティングの手法[Wilpon 92]の利用も検討する必要がある。

(iii) 大語彙音声認識

マルチモーダルでの音声入力のタッチパネル等の簡易入力手段に対する優位性は、語彙数が少なくとも50単語を越える必要がある。また、語彙数が数百単語以上になれば、音声入力の適用領域は、株価案内、列車や飛行機の情報案内、電話番号案内などに飛躍的に広がる。また、多少の認識率の劣化があっても、情報案内タスクでは、データベースによって、入力の曖昧性を減らすことも可能であるし、マルチメディア出力では、検索内容をCRTに表示することによっても認識結果の曖昧さや認識誤りに対処することができる。いずれにしろ、音声認識の精度をあげることが重要な課題である。話者適応によって精度をあげることや、不特定話者用に音韻コンテキストを考慮したHMM音韻モデルを作成するための多量の音声データベースの収集も不可欠である。

(iv) 発声方法

発声方法として、孤立単語発声、連続単語発声、自由発声の3種類が考えられる。タスクの認識項目が単純なANSER、窓口業務案内、列車の座席予約、株価問い合わせ、電話番号案内等は孤立単語発声モードでも可能と考えられる。これらのうち、認識の項目が多い列車の座席

予約、電話番号案内等は、連続単語発声モードも許す方が、音声インタフェースとして優れている。地理案内、ATIS、旅行相談等のように、認識の項目の構造がやや複雑なものは、自由発声モードを受け付けることが望ましい。しかしながら、自由発声モードの場合は、どのような構文表現まで許すのか、その許容範囲をユーザに明示することが難しい。音声入力のメリットを活かすには、少なくとも、連続単語発声を許容し、連続単語発声モードでも、多少の不要語や、決まり文句は許容できるようにすることが望ましい。また、認識誤りが起こるとユーザは一音節ずつ区切って発声する等の非常にゆっくりした発声になることが多く見受けられる。これを避けるために、ユーザに通常の発声をするように指示することも有用である。発声方法について、最初に簡潔に説明を与える方法も検討する必要がある。

（v）言語処理

音声認識での言語処理の役割は、言語情報により次に認識すべき単語や音韻をできるだけ限定することである。このような観点に合致する単純な言語モデルとして、有限オートマトンや単語や音節や文字のトライグラムが用いられている。この言語モデルの予測能力の評価に、パープレキシティという情報量的な尺度が用いられている。このようなトライグラム的アプローチは、モデルは単純であるが、単純であるために大量のテキストデータベースを利用できる点で大きなメリットを持っている。しかしながら、このような言語モデルの骨組みとして、有限オートマトンやトライグラムでなくて、文脈自由文法を用いようとする研究も増えてきている。特に、一般化予測ＬＲパーザは広く使われ始めている。さらに、チャートパーザやユニフィケーション技術[Okada 91]などを用いようとする小規模な試みもおこなわれている。

また、自由発声に対処するために、入力音声全体がパーズできなくても、発声内容の一部だけでも解釈できるようにパーザを動作させる部分パージングの手法も検討されている。このように自由発声を取り扱うためには、発声内容を理解することが重要で、一音一音正しく認識することは、必ずしも必要ではない。特に、情報案内タスクのように意味の構造（スロットの構造）が単純なものは、キーワードの認識によって意味内容の把握が可能である。以下、このような意味内容を考えた音声認識のための統計的言語モデルの研究課題について次の事項があげられる。（１）意味を基本としたパープレキシティの提案、（２）意味を考慮した統計的言語モデル、（３）文脈自由文法にユニフィケーションなどを組み込んだ効率のよい単語予測言語モデル、（４）対話状態に応じた言語モデルの適応化、（５）人と機械との対話を想定した自由発声テキストデータベースの収集とannotation。

システムとしては、連続単語発声と指示してあっても、単純な不要語や助詞の挿入や、文頭と文末の単純な表現は認識できるようにしておくことが有効と考えられる。

（vi）連続音声認識アルゴリズム

語彙数や構文が複雑になり、かつ不特定話者の音声に対処するためには、連続音声認識アルゴリズムも大幅に改良する必要がある。この改良には、音声認識精度の向上と認識速度の向上とがある。まず、認識精度の向上に関しては、ビームサーチでの枝刈りが適切におこなわれるように、フォワードだけの音韻モデルのスコアだけでなく、バックワードのスコアや数音韻先までのスコアを考慮することが必要となる[Minami 92][南 92]。また、言語モデルが文脈自由文法などになった場合の候補列のマージ（候補の融合）のアルゴリズムを検討する必要がある。特に、自由発声の認識の場合は、同じ意味を持つものが上位に並び、実際には、N－bestアルゴリズムが有効に動作しないことが知られている。意味を考慮したN－bestアルゴリズムを検討する必要がある[南 93b]。

また、自由発声入力の場合は、計算時間の削減も不可決となる。特に、トレリスの計算範囲を適切に制御し、計算量を減らすことが重要である。また、HMMの継続時間長制御による計算量の増加に対処することを考える必要がある。

（vii）対話モデル

音声は基本的に対話の手段としては、人にとって最も使い易いメディアである。一方、コンピュータにとって得意とするメディアは、音声応答でなく、むしろ、画像や図形の視覚情報である[市川 92][竹林 92]。音声応答は、その伝送情報量が少ないために、あまり、コンピュータ出力として有効ではない。むしろ、人に発話を促したりする、いわゆる割り込みに用いることが有用と考えられる。また、情報案内タスクは、旅行案内やATISを除くと、比較的単純な情報の検索タスクであり、複雑な対話モデルを必要としない。よって、情報案内タスクでは、コンピュータのメディアの特質を考慮した対話モデルを構築することが重要と考えられる。例えば、コンピュータによる認識結果を表示して、認識誤りをユーザにわからせること、できるだけ認識結果の確認をおこなわずにN－best候補からの検索結果を表示すること、音声応答は簡潔にすることなどが考えられる。その他、コンピュータとの対話の結果を常に収集して、対話テキストデータベースを集めることも重要である[速水 91]。

さらに、人と人との対話と簡単に比較してみる。人と人との対話は、相づちとか、不要語などが多く含まれており、非常に密な情報のやりとりがおこなわれる。人と機械との対話はどうであろうか。現在の音声認識の技術レベルでは、人ほど正確に、かつ、すばやく音声の認識を行なうことは、望むべくもない。よって、おのずから人と機械の対話は、人と人との対話とは異なってくる。このように、人と機械との対話という観点でシステムを作成することが非常に重要となってくる。

(viii) システムの評価

対話システムを評価する方法や評価尺度を見いだすことは、システム開発にとって重要である。この評価尺度として、タスクの完了率、タスクの達成までの所要時間、コンピュータとのやり取りの回数、キーワードの認識率などが調べられている[Seneff 91][Rudnicky90][Shriberg 92]。応答時間、発声の自由度、音声入力のスイッチのあるなし、音声応答中の入力（barge-in）の可否、雑音への耐性など様々な要因がシステム評価に関係する。

システムの典型的な評価方法は次のようである。まず、ユーザに簡単な説明と音声入力の操作を説明する。ついで、情報案内の達成目標（シナリオ）を与え、ユーザにシステムを利用してもらう。途中の音声入力、コンピュータの出力をすべて記録しておく[Hirshman 92][Pao 92]。最後にアンケートに答えてもらう。このような実験では、ユーザのシステムへの適応性、ユーザの目的達成意識なども重要なファクタであり、ユーザの利用記録を残しておく必要がある。

3.4.3 情報案内タスクの例

ここでは、将来の情報検索の分野で研究が開始されている大語彙連続音声認識技術を必要とするDARPAプロジェクトでおこなわれている航空機のフライト情報案内、いわゆる、ATIS（Air Travel Information System）[Austin 91][Ward 91a][Ward 91b]と、NTTの研究所で研究が開始されている電話番号案内サービス[南 92][鹿野 92]を取り上げる。ATISのシステム構成については、文献[ICOT-JIPDEC 91]のpp.75-83ですでに解説をおこなったので、ここでは、システムの評価の一例を紹介する。また、音声入力による番号案内は、現在の大語彙連続音声認識技術や不特定話者音声認識技術、自由発声の認識技術の問題点をより明確にするために説明を加える。

(1) ATIS（Air Traffic Information System）

ここでは、音声対話システムの評価方法の一例として SRIのATISシステム[Murveit 91][Jackson 91]を用いて行なわれた実験を紹介する。その前に、SRIのATISシステムの概要を説明する。

SRI ではDARPAのSpoken Language System (SLS)のプロジェクトの一環として、ATISタスクを取り上げて研究をおこなっている。ATISタスクは、従来のDARPAの不特定話者連続音声認識のResource Managementタスク (RM) よりも、音響処理においても次の点で挑戦的な課題である。

(i) 実使用環境下での音声（周囲雑音）

(ii) 未登録語への対処(out of vacabulary)

(iii) 自由発話の取り扱い(spontenuous speech)

特に、自由発話は、RMタスクで扱ってきた読み上げ文(read speech)と大幅に異なっている。例えば、自由発話の中には、言い直し(false start)、不要語(non-word)を含んでおり、かつ、調音も不明瞭になりがちである。さらに、自然言語処理の研究としても、きちっと書かれた文ではなく、話し言葉の解析をおこなわなければならないので、非常に挑戦的な課題である。

SRIのATISシステムは、単語列の認識を行なう音響処理DECIPHERと、誤りを含んだ単語列をも扱いうる解析および応答部の言語処理Template Matcherとに分けられる。

(a) 音響処理（DECIPHER) [Murveit 91]

DECIPHERの入力は、FFTベースのメルケプストラムとデルタケプストラムである。また、単語の発声のゆらぎを考慮した音韻ネットワークを作成するために、調音規則(phonological rule)を用いている。ATISのタスクでは平均で一単語あたり75通りの発音記号列を生成している。音韻HMMとして、環境独立の音素モデルからTriphoneまでを用い、音韻ネットワークでの頻度に応じて音韻HMMを選んで作成する。

(b) 言語処理(Template Matcher) [Jackson 91]

自然言語処理の研究者にとってATISタスクのような簡単に思われるタスクでも、話し言葉入力では、システムが予期できない構文が現われることがわかった。この言語処理 (Template Matcher) では、単語列全体を解釈できなくても意味内容をある程度把握できること(partial parsing)、解釈のあいまいさに対してはスコアを与えれることなどのメカニズムを考えている。

このシステムは、20数Mipsのワークステーション上にインプリメントされており、実時間の数倍の時間で応答している。また、このシステムのハードウェアバージョンも作成されている。

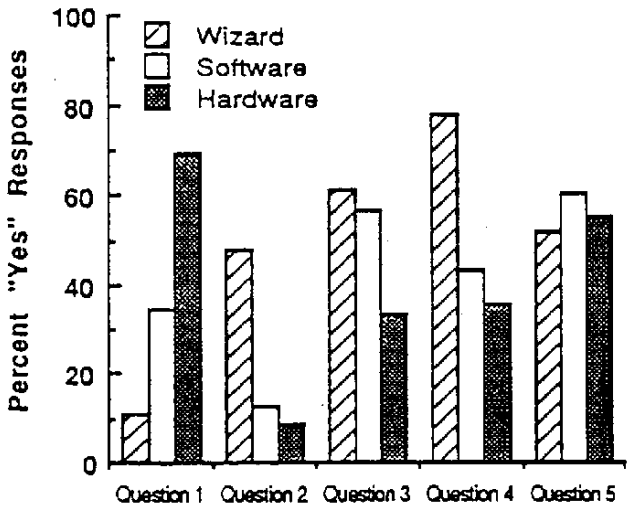
このATISのシステムを用いて、ユーザの満足度とシステムの性能の関連を文献[Shriberg 92]では調べている。以下、これについて紹介する。この研究では、次の3つの事柄について、実際にユーザを用いて検討している。(i) 認識速度と認識率、(ii) ユーザのシステムへの適応化、(iii) 誤りが起こった場合のユーザの発声。ユーザへのインタフェースとしては、マウスとマイクロフォンが用意されているが、キーボードは隠してあり、利用できない。音声入力はマウスのクリックで開始される。クリックの0.5秒前から音声が取り込まれ、2秒以上の無音が続くと終わりと見做され、音声の取り込みが終了する。認識結果の文章と、結果をまとめた表がCRTに表示される。また、対話の制御として、マウスのクリックで前のシステム状態に戻ることもできる。ユーザは、使う前に、簡単なデモによって、使い方を学習する。さらに、

半分のユーザは音声の発声方法などの指示をインストラクタから受けるようにしている。

上記の(i)～(iii)の事柄を調べるために次の3種類のシステムを用意した。(A) 速度は遅い（2,3分）けれどほとんど完全な認識結果（Wizard方式）、(B) 精密なソフトウェアバージョン（単語誤り率16.1%）で実時間の数倍で動作、(C) ハードウェアバージョンでほとんど実時間で動作するが、認識率が若干低いもの（単語誤り率24.8%）。システムのユーザは145人で、約1／3ずつ、それぞれのシステムを利用した。実行するシナリオとして、次の4種類を用意した。(s1) 3日の内に2つの市を訪れる。(s2) ファーストクラスの異なった料金を見つける。(s3) 3箇所からのフライトの到着時間を調整する。(s4) 2つのフライトのうちから料金と食事とを考えて選択する。

これらのシナリオのタスクを終了した後、図3.4-2の下に示すようなアンケートを取っている。質問1の応答速度では、当然のことながら(C)のハードウェアが満足の領域に達している。しかしながら、質問2の、「システムがあなたの質問を理解しているか？」に対しては、(A)のWizardシステムのみが満足レベルに達している。さらに、質問4の、「計算機を余り知らないユーザでもこのシステムを使えると思うか？」に対しては、認識率の高いシステムほど、高い評価を得ている。質問5の、「フライト表をひくよりこのシステムのほうが好ましいか？」に関しては、半数以上の人が、いずれのシステムに対しても肯定的であった。

次に、(ii)のユーザのシステムへの適応化に関しても上記の実験結果をもとに考察している。シナリオ(s1)と(s2)とでのユーザの入力を分析している。(s1)で単語認識率が低かったユーザは、システムに対して、決まり切った単純な構文や簡単な単語を使うようになり、(s2)での単語認識率が著しく向上していることが分析されている。特に、簡単な構文によって、言語モデルのパープレキシティの低いものを使うようになることが述べられている。



Question 1: Were the answers provided quickly enough?

Question 2: Did the system understand your requests the first time?

Question 3: I focused most of my attention on solving the problems, rather than trying to make the system understand me.

Question 4: Do you think a person unfamiliar with computers could use the system easily?

Question 5: Would you perfer this method to looking up the information in a book?

図3.4-2 ユーザの満足度［Shriberg 92］

最後に、(iii)の誤りが起こった場合のユーザの発声についても考察している。このシステムの音韻モデルは、Wizardシステムで集められた自由発声をもとに作成されている。ユーザは誤りが生じるとゆっくりと丁寧に発声し、幾人かのユーザでは、誤りを増やすことになっている。始めに発声方法について指示を与えた半分のユーザのほうが若干高い単語認識率を示していた。ただし、自由に発声してください(Speak naturally)という指示は、不要語や言い直しやラフな発声を引き起こし、適切な指示ではなかったと述べられている。
このように、音声対話システムを評価する研究が、米国のDARPAプロジェクトを中心におこなわれ始めている。今後は特に、マルチモーダル、マルチメディアのパラダイムでの評価が重要となってくるものと考えられる。

（2）　音声入力による番号案内

自由に発声された音声を認識するためのシステムの研究が始められている。ここでは、電話番号案内を取り上げ、自由発声の認識の試みを文献[鹿野 92]に基づいて紹介する。この電話番号案内のタスクには、東京都の三鷹市と武蔵野市の全個人加入者の住所と名前が含まれている。よって、加入者総数、約74,000人の名前の認識をおこなうことになる。このような大語彙を含んだタスクでの自由発声でも音声認識ができることの可能性を示している。このように大語彙を含む対象でも音声認識が可能であれば、メディアとしての音声入力の適用領域の可能性は、飛躍的に広がるものと思われる。

ここで取り上げられた電話番号案内は、目的がはっきりしており、認識すべき項目（スロット）も明確である。スロットとして、表3.4-2に示す5項目がある。それぞれの項目の異なり単語数も表3.4-2に示しておく。

表3.4-2　電話番号問い合わせタスク

電話番号問い合わせタスク （東京都の武蔵野市と三鷹市の個人加入者）	
加入者数	73、693名 （武蔵野 34,143, 三鷹 39,550）
都	1通り　（東京）
市	2通り　（武蔵野、三鷹）
町名	27通り
番地	620通り（×-○○）
名前	71、251通り （名前 62,659、姓のみ 8,592）

電話番号の問い合わせでは、人が必ずしも町名や番地を正しく発声しないことが多いので、電話帳にある項目間の制約は用いていない。認識実験のための自由発声文を、電話番号案内の模擬対話書き起こし文から作成している。人と人との対話では相づちなどが含まれているが、自由発声文の作成では取り除いている。また、氏名の漢字の説明など、著しく複雑な文（全体の約30%程度）は、除外している。認識実験に用いた自由発声文の例を連続単語発声の例とともに表3.4-3に示しておく。

表3.4-3　連続単語発声文と自由発声文の例

自由発声文	（1）あ、ええと、すみません、ええっと、武蔵野市の境南町のですね、××さん、お願いします。
	（2）三鷹市内で、住所は、三鷹市、新川×丁目○○にお住まいの、××○子さんって方の電話番号。
連続単語発声	（1）武蔵野市境南町、××さん。
	（2）三鷹市、新川×丁目○○、××○子さん。

自由発声には、不要語や様々な言い回しや不完全文などが含まれていて、構文の記述が困難であるとされている。このための1つの手法として、ワードスポッティングが、ガーベージモデルや未知語モデルなどとともに提案されている。しかしながら、単語数が数千のタスクになると、キーワードのスポッティングは不可能となる。未知語の検出などはまったく不可能となる。

ここでのタスクは、語彙数が多いけれども、目的がはっきりしており、スロット構造も単純である。よって、多くの発声例を参考にして、文脈自由文法である程度記述できるとしている。また、文法から若干はずれた文でも、音声認識のあいまい性のために、類似の構文に認識され、キーワードが正しく認識される可能性も大いにある。文法の記述を図3.4-3に示すように、階層的に行なっている。各スロットの構文を、電話帳のデータベースを書き換え規則の形に変換することによって作成している。これらのスロットの構文を制御する文形（メインシンタックス）は、自由発声の例文（約300）を参考にして作成している。これらの文法を図3.4-3のような2段階からなる一般化予測LRパーザによって起動している。LRパーザは、直接にHMM音韻モデルを起動する方式になっている。例文から得られた音韻系列をLRパーザに入力

して、アクセプトされるようにチェックしながら、書き換え規則の記述を行なっている。約半分の音韻系列は一意にパーズされている。残りの半数は、2通りのパーズがなされている。これは、表3.4-3の文例2のような市の名前を2回発声しているものをアクセプトするために、市の名前のスロットを2つ設けたためである。表3.4-4にそれぞれのスロットの書き換え規則数とLRテーブルの状態数を示しておく。

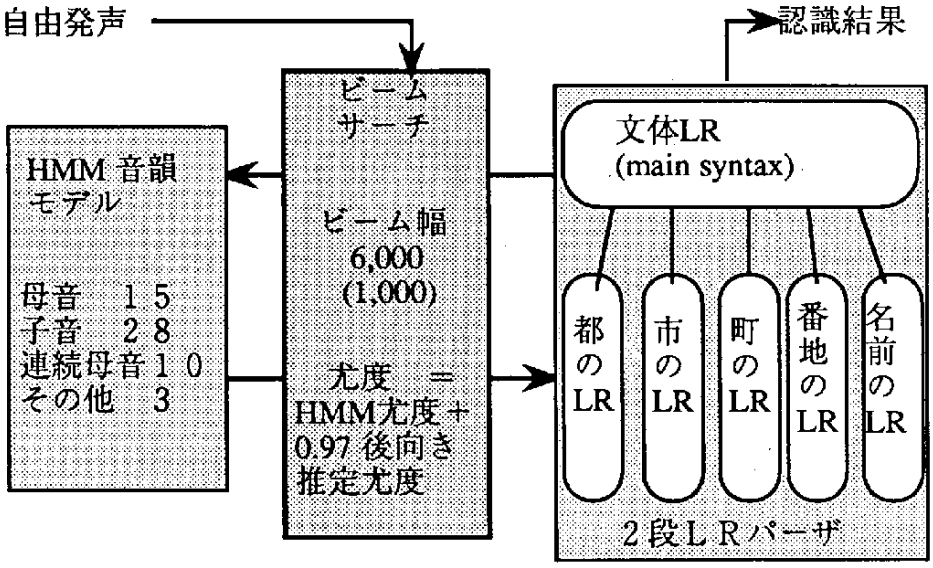


図3.4-3　電話番号安寧をタスクとする自由発声認識システム

表3.4-4　各項目の構文と複雑さ

		文体	都	市	町	番地	名前
自由発声文	書き換え規則数	470	5	9	48	1,239	71,262
	LRテーブルの状態数	1,069	9	24	217	2,756	312,025
連続単語発声	書き換え規則数	15	5	9	48	1,239	71,260
	LRテーブルの状態数	22	9	24	217	2,756	312,016

自由発声と単語発声の構文の複雑さを評価するために、テスト文に対する音韻パープレキシティを求めている。その結果を表3.4-5の音韻パープレキシティの項に示しておく。また、比較のために日本語の音節連鎖の制約だけを用いた場合の音韻パープレキシティの表3.4-5の最下段に示しておく。自由発声の構文のパープレキシティをみると、連続単語発声文に対するパープレキシティの方が自由発声文に対するものより大きくなっている。これは、不要語や文末の言い回しなどの冗長性がパープレキシティを下げる働きをしているためであるとされている。これらのことは、自由発声文の認識の困難さを、必ずしもパープレキシティで表すことができないことを示唆している。

表3.4-5　認識実験結果とパープレキシティ

		発話モード			連続単語発声	自由発声
音韻認識率		第1位の認識率			72.9 %	69.1 %
		2位以内に含まれる率			88.1 %	84.1 %
		5位以内に含まれる率			98.0 %	95.2 %
文理解率	構文の種類	連続単語発声の構文		音韻パープレキシティ	4.09	-
			ビーム幅1000	文理解率	88 %	-
				項目認識率	97 %	-
		自由発声の構文		音韻パープレキシティ	7.58	6.39
			ビーム幅1000	文理解率	78 %	33 %
				項目認識率	92 %	67 %
			ビーム幅6000	文理解率	80 %	59 %
				項目認識率	95 %	84 %
日本語の音節連鎖の制約				音韻パープレキシティ	25.25	24.77

男性1名（ec10002)が発声した連続単語発声と自由発声を用いて、特定話者音声認識実験をおこなっている。HMM音韻モデルは、200文のATR音韻バランス文の発声データから学習している。音韻の基本単位（ユニット）として、図3.4-3に示す56音韻（ユニット）を用いている。連続単語発声と自由発声に対する音韻（ユニット）認識率を表3.4-5に示しておく。

連続単語発声は、連続単語発声の構文でも自由発声の構文でもアクセプトされるので、両方

の構文で認識実験を行なったている。自由発声文は連続単語発声の構文ではアクセプトされないので、自由発声の構文のみで認識実験を行なっている。自由発声の認識では、パープレキシティの増加に対処するためにビーム幅（音韻系列とLRの状態の途中結果の候補数）を6,000に設定している。

自由発声に対して文理解率59％、項目の認識率84％が得られた。連続単語発声に対しては、大きいパープレキシティにもかかわらず、高い文理解率80％と項目の認識率95％が得られている。さらに、連続単語発声に対しては、ビーム幅を1,000に設定しても認識率の低下は少ないことも分かる。また、連続単語発声の構文で連続単語音声入力を認識した場合には、文理解率88％、項目認識率97％であった。この結果は、発声が連続単語入力であれば、自由発声の構文でも高い認識率が得られることを示唆している。よって、ユーザには連続単語入力を指定しておいて、認識システムの構文では自由発声を受け付けるように設定しておく方式の可能性を示唆している。

自由発声入力での誤りの多くは、不要語や市の名前を人の名前（例、anoo -> ano, mitakashi ->itabashi）に間違えることにより、途中結果のビームから正しい候補がはずれてしまうことに起因している。つまり、不要語や市の名前を人の名前に間違えた場合、人の名前の部分までその誤りは訂正できないという、構文表現のある種のlong distance dependency の問題によっている。さらに、自由発声での文頭表現や不要語や項目の語尾の言い回し等が、音韻系列のバリエーションを指数関数的に増大させている。これに対処する一方法として、音韻系列のマージを意味レベルでおこなうことが考えられる[南 93b]。その他、left-to-rightのアプローチは、上記のような long distance dependency に対して本質的に弱いという問題が生じている。

さらに、不特定話者の音韻HMMを用いて連続発声や自由発声を認識しようとする試みも行なわれている[南 93b]。また、発声環境の周囲雑音に音韻HMMを適応させる研究も行なわれている[南 93a]。

このような番号案内での人と機械との対話について想像してみる。まず、最初の画面の音声入力ボタンを押して、音声入力をおこなう。システムは、上位候補から電話帳の住所、名前を調べて、図 3.4-4のようにCRT上に表示する。その表示された中に、ユーザが調べたい名前があれば、ユーザはその候補をマウスでクリックして、電話番号を知り、対話は終了する。そうでなければ、認識結果の誤った項目をマウスでクリックして、もう一度発声する。どうしても認識ができない場合には、リセットをクリックして最初の画面に戻るか、キーボードモードにしてキーボードから入力する。このようなストーリーで対話システムを構築することができる。

この例では、機械からの応答文が簡潔に表示されるし、また、認識結果も一番下に表示されている。認識結果の表示は、人が機械の認識能力（受け付ける構文、単語）を推察して、機械

に適応するのに重要な情報となると思われる。このようにマルチモーダル、マルチメディアのパラダイムでは、人と機械との対話は、人と人との対話と全く異なる形態をとる可能性が大きい。

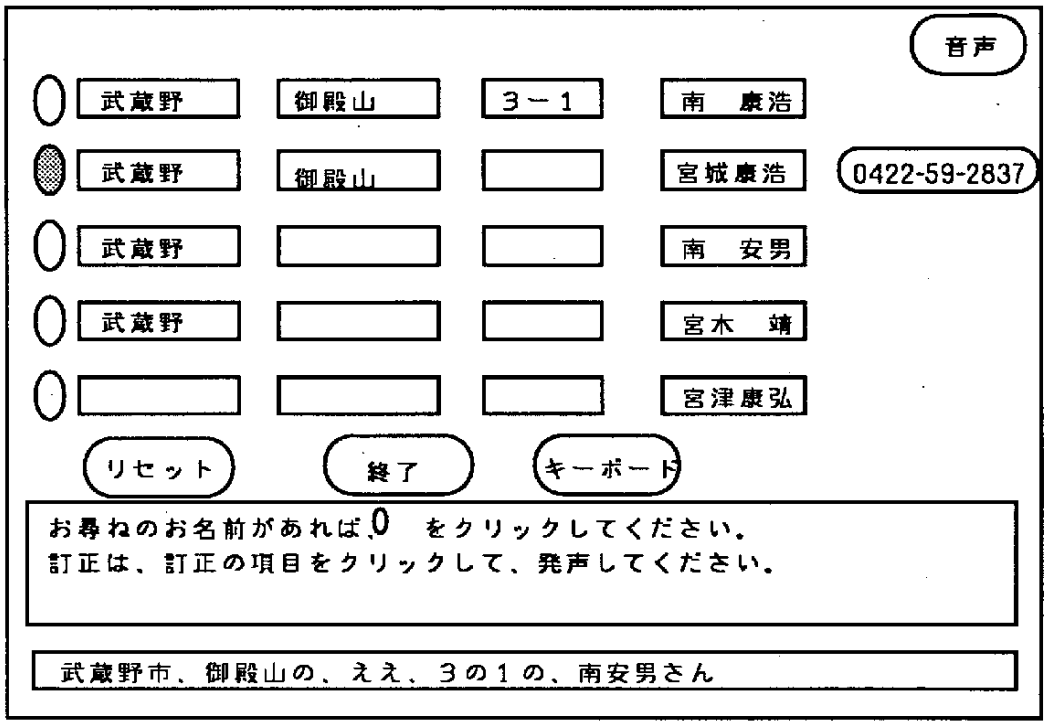


図3.4-4　電話番号案内での表示画面の例

3.4.4　まとめ

音声入力による情報案内システムについて、想定できる数種のタスクを取り上げ、さらに、音声認識を中心として技術課題について述べた。さらに、具体的な研究例として、DARPAのATIS（Air Travel Information System）とNTTの電話番号案内を取り上げた。ATISはシステムの評価例として、番号案内は大語彙連続音声認識の例として取り上げた。

ここでは、韻律情報の利用についてはあまり取り上げなかったが、自由発声の認識をおこなうには、不要語の抽出、キーワードの検出、文節境界の検出などに有効に利用できる可能性がある。また、情報の検索においても、プライバシーに関するものに対しては、セキュリティの問題も重要となり、暗証、話者照合、暗号化等の技術が不可決となる。

また、マルチメディアにおいて、機械を擬人化するのがよいのか、機械は機械らしくみせる

のがよいのかということも、これからの研究課題である。

音声による情報案内サービスは、ユーザのサービスの拡大と向上の要求に答え、かつ、人件費の削減（合理化）の可能性も持っており、非常に有望な分野である。今後、急速に普及して行くものと思われる。

参考文献

1.[Austin 91] S.Austin, et al., BBN HARC and DELPHI Results on the ATIS Benchmarks-February 1991, Proceedings of the DARPA Language Workshop, Morgan Kaufmann Publishers, Inc., pp.112-115, (1991-02)

2.[Billi 92] R.Billi, An Overview of Speech Synthesis and Recognition for Telephone Applications in Italy, 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Piscataway, New Jersey, (1992-10)

3.[Gales 92] M.J.Gales, S.Young, An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise, Proceedings of ICASSP90, pp.I-233-236, (1990-03)

4.[Harlow 92] J.Harlow, R.Rosinski, VALID: Voice Activated Licence Identification, 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Piscataway, New Jersey, (1992-10)

5.[速水 91] 速水　悟、伊藤克亘、田中和世、音声対話システムの構築とそれを用いた会話音声収集、電子情報通信学会音声研究会資料、ＳＰ－９１－１０１、(1991-12)

6.[Hirschman 92] L.Hirschman, Multi-Site Data Collection for a Spoken Language Corpus, MACOW, Proceedings of the DARPA Language Workshop, Morgan Kaufmann Publishers, Inc., pp.7-14, (1992-02)

7.[市川 92] 市川、天野、北原、口語調音声による情報検索システムの試み、科研費総合A「音声対話のモデル化とその機械処理に関する総合研究」最終報告書、pp.93-99, (1992-02)

8.[ICOT-JIPDEC 92] ICOT-JIPDEC AI センター、音声の知的処理に関する調査研究、(1992-06)

10.[金田　92] 金田豊、アダプティブマイクロホンアレー、電子情報通信学会論文誌　B-II, No.11, pp.742-748, (1992-11)

11.[Lennig 92a] M.Lennig, et.al., automated Bilingual Directory Assistance Trial in Bell Canada, 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Piscataway, New Jersey, (1992-10)

12. [Lennig 92b] M.Lennig, D.Sharp, V.Gupta, P.Kenny, K.Precoda, Flexible Vocabulary Recognition of Speech Over the Telephone, 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Piscataway, New Jersey, (1992-10)
13. [Martin 92] F.Martin, K.Shikano, Y.Minami, Y.Okabe, Recognition of Noisy Speech by Composition of Hidden Markov Models, 電子情報通信学会音声研究会資料、ＳＰ－９２－９６、(1992-12)
14. [Minami 92] Y.Minami, K.Shikano, T.Yamada, T.Matsuoka, Very Large Vocabulary Continuous Speech Recognition for Telephone Directory Assistance, 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Piscataway, New Jersey, (1992-10)
15. [南 92] 南、山田、鹿野、番号案内を対象とした大語彙連続音声認識アルゴリズム、電子情報通信学会音声研究会資料、ＳＰ－９２－１０８、(1992-12)
16. [南 93a] 南、マルタン、鹿野、HMMの合成による雑音環境下の大語彙連続音声認識、日本音響学会春季研究発表会講演論文集、(1993-03)
17. [南 93b] 南、山田、吉岡、鹿野、自由発声音声認識における意味を考慮した２段LRパーザ、日本音響学会春季研究発表会講演論文集、(1993-03)
18. [Murakami 87] K.Murakami, et al., Voice Inquiry System in a Public Telephone Network, Globcom87, pp.1187-1191, (1987]
19. [Murveit 91] H.Murveit, J.Butzberger, M.Weintraub, Speech Recognition in SRI's Resource Management and ATIS System, Proceedings of the DARPA Language Workshop, Morgan Kaufmann Publishers, Inc., pp.94-100, (1991-02)
20. [NTT 92] NTTデータ通信、ANSERシステム（パンフレット）
21. [Okada 91] M.Okada, a Unification-Grammar-Directed One-Pass Search Algorithm for Parsing Spoken Language, Proceedings of ICASSP91, S10.9, pp.721-724, (1990-03)
22. [Pao 92] C.Pao, J.Wilpon, Spontaneous Speech Collection for the ATIS Domain with an Aural User Feedback Paradigm, Proceedings of Speech and Natural Language Workshop, Morgan Kaufmann Publishers, Inc., pp.43-48, (1992-02)
23. [Rosenbeck 92] P.Rosenbeck, Baungaard, A Real-World Telephone Application: teleDialogue Experience and assessment, 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Piscataway, New Jersey, (1992-10)
24. [Rudnicky 90] A.Rudnicky, The Design of Spoken Language Interface, CMU Technical Report, CMU-CS-90-118, (1990-03)

25. [Seneff 91] S.Senef, et al., Development and Preliminary Evaluation of the MIT ATIS System, Proceedings of the DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, Inc., pp.88-93, (1991-02)

26. [鹿野 92] 鹿野、南、山田、番号案内を対象とした自由発声の認識の試み、日本音響学会春季研究発表会講演論文集、1-1-7、pp.33-34、(1992-10)

27. [島田 90] 島田、牧野、最近の電話の音響技術、エコー制御技術、テレビジョン学会誌、44, 3, pp.222-227, (1990-03)

28. [Shriberg 92] E.Shriberg, E.Wade, P.Price、Human-Machine Problem solving Using spoken Language System (SLS) : Factors Affecting Performance and User satisfaction, Proceedings of Speech and Natural Language Workshop,Morgan Kaufmann Publishers, Inc., pp.49-54, (1992-02)

29. [Sorin 92] C.Sorin, et.al., CNET Speech Recognition and Text-to-Speech in Telecommunications Applications, 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Piscataway, New Jersey, (1992-10)

30. [竹林 92] 竹林、坪井、音声による計算機との対話システムの試作について、電子情報通信学会第２種研究会資料、pp.7-8, (1992-02)

31. [武田 92] 武田、小沼、自由発話文理解にためのGarbage HMMの利用の検討、電子情報通信学会音声研究会資料、ＳＰ－９２－１２７、(1993-01)

32. [Varga 90] A.P.Varga, R.K.Moore, Hidden Markov Model Decomposition of Speech and Noise, Proceedings of ICASSP90, pp.845-848, (1990-03)

33. [Voice 92] 1st IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Piscataway, New Jersey, (1992-10)

34. [Ward 91a] W.Ward, Understanding Spontaneous Speech: The phoenix system, Proceedings of ICASSP91, S5.29, pp.365-368, (1991-05)

35. [Ward 91b] W.Ward, Evaluation of the CMU ATIS System, Proceedings of the DARPA Language Workshop, Morgan Kaufmann Publishers, Inc., pp101-105, (1991-02)

36. [Wilpon 92] J.G.wilpon, D.B.Roe, AT&T Telephone Network Applications of Speech Recognition, Workshop on Speech Recognition Over the Telephone Line, (1992-11)

3.5 予約システム

3.5.1 背 景

「チケットぴあ」に代表されるようなチケット予約のオンライン化の発達によってチケットの予約が電話一本でおこなえるようになってきた。しかし、予約センターの電話はチケット発売開始とともにずっと通話中になり、なかなか電話がつながらないという不満は、このようなシステムを利用したことがある人なら誰しも経験したことがあるだろう。本節では、近い将来利用可能になるであろうコンピュータ技術によって現行の予約システムがどのように改善されうるかについて、その利用形態、技術的課題という側面から考察する。

3.5.2 予約システムの利用形態

（1）現行の予約システムの問題点と改善

予約センターに電話をしてチケットを予約しようとしたが、電話が話し中でぜんぜん電話がつながらなくてイライラしたという経験は誰しも持っているだろう。チケット予約の開始時期には特にこの傾向がひどく、ダイヤルしては通話中のトーンを確認して、すぐ電話を切り、またすぐにダイヤルするといったことを繰り返すはめになる。要するに利用者の数に対して電話回線の数が圧倒的に少ないためにこのようなことが起るわけであるが、電話回線を臨時に増やすことができても、その電話回線の向こうにいるオペレータの数を柔軟に調節することは、それほど容易ではない。また、コストの面から考えてもオペレータの人件費は無視できないだろう。もし、人間のオペレータの代りにコンピュータが予約業務を代行できれば、必要な回線の数を柔軟に変化させることがもっと容易になるであろう。

コンピュータに予約業務を代行させることを考えるときに、もっとも重要となる技術は利用者の意図を正しく認識することである。自然言語処理の研究分野では対話に関する研究がいくつかおこなわれている［Winograd 72］、［Bobrow 77］。対話では、対話参加者がお互い相手の知識をうまく利用して、相手が知っていることは極力明示的に表現しないなどの方略をとって、対話をできるだけ効率よく進めるのが普通である。つまり、人間が持ってる常識や対話の話題となっている領域の知識をうまく使って対話をおこなうのである。しかしながらコンピュータに常識を持たせようという研究は現在もおこなわれているが［Lenat 86］、これを実現するためにはまだしばらく時間がかかると思われる。これまでの対話研究では、対象領域を非常に限定し領域知識をできるだけ詳細に記述するとともに、常識はその領域に必要な最小限のものを用意するという方法によりある程度の成果をおさめてきた。予約業務という領域に限定し、不十分な点はエラー回復をどのようにおこなうかなど、ユーザインタフェースの技術でカバーすることによって現在の技術でも一定の成果は得られるだろう。

問題は利用者の発話をどの程度正しく認識できるかである。これまでの対話研究のほとんど

は利用者の入力はキーボードでおこなわることを仮定している。つまり、入力に関しては利用者の意図した発話が正しく入力されることを仮定している。しかしながら音声による対話を考えた場合はこの仮定は成り立たなくなる。特に予約業務という性質を考えると、システムのエラーが利用者に金銭的な被害をもたらす可能性があるので、信頼性、セキュリティの点でも慎重になる必要がある。この点で、システムが情報を与えればそれでよいという情報案内システム（3.4節参照）とは性質が異なるのである。音声対話の研究はまだ始まったばかりであり［長尾 93］、［伊藤 92］、電話による予約業務をコンピュータが代行できるようになるには、さらに研究が必要だろう。

（2）予約システムの高機能化

現在の電話によるチケットの予約業務をコンピュータに代行させるためには音声対話の技術の確立が必要であることについて述べたが、もうひとつの方向性として、現在の予約業務では利用者に提供されていないようなサービスの提供を考えることができよう。劇場に予約システムでチケットを予約することを考えてみよう。予約センターへ電話をし(運がよければ電話がつながって)、自分の希望するチケット、日時、席の種類などをオペレータに告げることによって予約が完了する。希望する席を具体的に指定できる場合もあるが、これは実際にチケットブースに出むいて、その場で確認する場合に限られることが多く、電話でこれを実現することはむずかしい。実際には希望する座席をこまかく指定できるサービスの需要は多いだろうが、現実には座席のクラスが指定できる程度に限られている。これは電話による予約では、情報伝達が音というモーダルに限られているからである。

たとえば、東京文化会館でおこなわれるコンサートの席を予約する場合を考えてみよう。自分が希望する席がどの位置であるかは、東京文化会館のホールの構造をよく知っている人ならホールの図面を見るだけで席の番号を知ることができるかもしれない。しかし、自分の希望する座席の番号がわかったからといって、その番号が必ずしも取れるとは限らない。もし、その番号の席が予約済みだったら、他の席を指定することになる。しかし、これを延々と繰り返していたらいつまでたっても席が決まらないばかりか、貴重な電話回線を占有してしまうはめになる。では、オペレータにまず空いている席を教えてもらったらどうだろうか。空席が少ない場合はこれも可能だろうが、席が何十とある場合はこれも実現が困難である。この問題は、オペレータと利用者が共有できる座席表を導入することによって解決できる。利用者が実際にチケットブースに出むき、チケットを購入する場合は、オペレータと利用者が目の前にある座席表を指差しながら席を決めることができる。座席が指定したければチケットブースまで来いというのでは、質の高いサービスとはいえないだろう。利用者がどこにいても同じサービスが受けられることが望ましい。そういう意味では、3.7節で述べる携帯端末からも同等のサービス

が受けられるようになるとよいだろう。

オペレータと利用者の間の情報交換のモダリティの拡大という意味では、自分の希望するコンサートのホールの構造がどのようなものであるかに関する情報についても、利用者が席を決める際に提示できるとよいだろう。このための情報として現在利用可能なのはホールの座席表を平面的に図示してものだけである。しかも、電話予約の場合は手元にホールの情報を自分の責任で用意しなければならない。そのホールに行ったことのある人ならば座席表の情報からだけでも自分の希望する席をある程度決めることはできるだろうが、一般に行ったことのないホールの座席の平面図だけを見て、自分がどこに座りたいかわかる人はいないだろう。このために最近研究が盛んな仮想現実感の技術（2.2節参照）を用いてその座席位置からホールがどのように見えるかをシミュレーションして、利用者に提示することによって利用者の希望にあう座席を探すことが可能になるだろう。

以上まとめると、今後の利用形態として、以下のような方向が考えられる。

1.音声ベースの予約システムの改善
2.移動性、携帯性の重視
3.多様なモダリティの導入

もちろんこれらは適当に組合せることが可能であり、互いに排反なものではない。たとえば、音と映像が扱える携帯端末が利用可能になれば、マルチモーダルなインタフェースを持ち、しかも、いつでもどこでも予約が可能なシステムの実現も可能となろう。次項では、これらの方向で予約システムを進化させるために必要な要素技術、およびその技術的課題について述べる。

3.5.3 技術課題

（1）対話処理

3.5.2項で述べたように、これまでの対話処理に関する研究では、領域を限定することによって問題を簡単にし、対話をうまく扱ってきた。しかし、領域が限られているとはいっても、人間同士の対話がうまくいくことの背景には人間の持つ常識に負うところが大きい。研究課題のひとつは対話処理に有効な人間の持つ常識にはどのようなものがあるかを明らかにし、それをコンピュータ上に実現する手法を考えることであろう。

常識の問題とも関係するが、もうひとつの研究課題として意図理解の問題があげられる。人間でも相手を誤解することがあるように、相手の意図を正確に理解することは人間にとっても困難な場合がある。コンピュータにとってはなおさらである。意図の理解は特に間接発話行為のように明示的な言語表現をともなわない場合は特に困難である。たとえば、部屋が暑いので

窓を開ける、あるいはエアコンのスイッチを入れて欲しいと思うときに、相手に向かって「窓を開けてください」あるいは「エアコンを入れてください」と発話した場合、聞き手はこれが依頼であることは発話の表層的な特徴から認識することができる。ある人が依頼をするとき、その人の意図はその依頼の内容、この例では、「窓が開く」あるいは「エアコンのスイッチが入る」が真になることである、という知識があれば、この発話の意図を理解し、適切な対応をとることは可能である。しかし、同じ意図を持っているときに「暑いですね」とか「窓を開けませんか?」という発話をすることも我々はできる。これらの発話は表層的には依頼を表してはいないが、発話者に意図は明らかに依頼である。たとえば、「暑いですね」と言われて「はい、そうですね」と答えるのはこの文脈では明らかに不自然である。しかし、道で偶然出逢って、天候のあいさつとして、「暑いですね」といわれて「そうですね」と答える場合にはまったく自然である。このようにその場の状況によって同じ発話でもその発話の意図はまったく異なる場合がある。領域をうまく限定することによってこのような問題は生じにくくなってくるが、より知的なシステムを構築するためには、意図の理解は避けて通れない問題である。

最後に音声との関係で、音声対話に関する研究も忘れてはならない。音声による対話では、いい間違いや不要語(「あー」とか「えーと」など)をどのように扱うかという問題が生じる。また、音声認識の技術にも限界があるので、意味処理あるいは対話処理の技術をうまく利用して、音声認識の誤りを補償する技術が必要である。音声認識の精度を向上させるために意味処理、対話処理の技術をどのように適用するかについては現在も盛んに研究がおこなわれている。

（2）マルチモーダル化

現行のチケットの予約の形態がほとんど電話を通じておこなわれ、3.5.2項で述べたような問題があることを考えると、インタフェースのマルチモーダル化は必要不可欠であろう。特に座席等の予約では、視覚的なインタフェースの導入が特に重要であると考えられる。究極的には仮想現実感の技術を利用して、その座席に座ったときの感覚をシミュレーションして利用者に擬似体験させ、利用者が席を選択するための情報として提供できるようになればよいだろう。しかし、2.2節でも述べたように、本格的な立体像を提供するためにはホログラフなどの技術が必要となるが、現在の技術水準ではこれを実現するのは困難である。より実現可能性の高い候補として、空席状況をオンラインで表示できる座席表を用意し、利用者がその座席表の座席を指差すことによって席を予約できるようなインタフェースが考えられる。コンサート名や日時は音声で指定し、 席は指で指示することによって指定するとういうことも考えられる。このような方法なら現在の技術をうまく利用すれば実現可能であろう。いずれにしても、 マルチモーダルなインタフェースを考えるときには、どのような情報をどのモダリティによって表現するかを適切に決定することが重要となる。そういう意味ではこれまですべて音声情報で伝

えていたものを適切なモダリティに移行させることによって、音声認識の負担は減ると考えられる。マルチモーダルなインタフェースの研究もまだはじまったばかりであり、これからの研究成果が期待される分野である。

（3）移動体通信

パームトップコンピュータやペン入力コンピュータが一部ではすでに実現されていることを考えると、将来的には携帯端末が普及することが予想できる。現在でも、移動電話はその料金が通常の固定電話に比べ桁違いに高いにもかかわらず、かなりの普及率をしめしている。携帯端末が普及すれば、いろいろなことを場所や時間の制約なしに携帯端末を通しておこなうことが可能になるだろう。チケットの予約もその例外ではない。現在ではチケットブースに出むくか、電話で予約するしかないので、平日の予約は仕事をさぼらない限り困難であるが、携帯端末から容易に予約が可能になれば、ちょっとした休憩時間を利用して予約をすることも可能になる。携帯端末には、文字情報だけでなく、音声やグラフィックスも扱う機能が必要である。

また、携帯端末がどこにあっても信頼性のある通信チャネルを確保するための通信ネットワークの技術も必要不可欠となる。移動体の移動にともなって起りうる死角をできるだけなくし、安定した通信を提供できなければならない。

（4）高信頼化、セキュリティ

3.4節の情報案内システムでは、利用者からの要求に対して、必要な情報をデータベース中から検索し、提供するのが主な機能であった。一方、予約システムでは、利用者の予約をデータベースに反映させなければならない。このためにデータベースに対する変更が正当なものであるという十分な保証がなければならない。不正なデータベースの変更はシステムに壊滅的な打撃を加える可能性もある。このためには、不正な利用者のアクセスを防止するためのセキュリティの技術が必要となる。現在、アクセス・コントロールはパスワードによっておこなうことが多いが、音声による話者認証もこのための候補となりうるだろう（2.3節参照）。また、正当な利用者による利用であっても、利用者の意図しないような変更がデータベースに加えられることも防がなければならない。特に音声を入力として用いる場合には認識誤りがないかを十分チェックする必要がある。このためには、利用者にシステムの認識した入力を確認させたりする必要がある。しかし、このような再確認はうまくやらないと利用者をうんざりさせる結果になる恐れがある。このためにはユーザインタフェースの研究成果や認知心理学の研究を参考にすべきである。

3.5.4 まとめ

本節では今後の技術・社会動向などを目途にして、予約システムを例にとりあげ、利用形態と技術課題という観点から考察した。まとめると以下のようになる。

1.想定される使用状況、利点、必要度

(a)パームトップのような携帯端末から移動体通信技術を用いて時間と場所を選ばずにチケット予約が可能になる。

(b)仮想現実感(virtual reality) の技術を用いた、よりきめの細かいサービスを受けることができるようになる。たとえば、座席を選択する際に、その座席位置からステージがどのように見えるかをシミュレーションすることによって、利用者の要求にあった座席を探すことが可能になる。

2.必要な機能、性能、装置の形態、構成

このようなシステムを実現するためには、帯域の広い移動体通信技術、軽くて小型の高性能コンピュータ、仮想現実感などの要素技術が必要となる。

3.これを実現するための技術要素、重要研究課題

要素技術	現在	5年後	10年後
言語	領域知識	意図理解	意図理解, 常識の実現
映像	CG	アニメーション	ホログラフ
通信	固定局	移動電話+携帯端末	移動体通信+携帯端末
音声	単語認識 100語彙	不特定話者 1000語彙	個人認証+話者適応 10000語彙

4.必要となる技術(キーワード)

(a)音声・言語関連分野

音声処理、韻律理解、対話処理、話者の意図の理解、音声合成、ハイパーテキスト、フルテキストサーチ

(b)通信ネットワーク技術

移動体通信（mobile computing）、ISDN、光通信

(c)関連分野

リアルタイムCG、アニメーション、表情合成、表情認識、仮想現実感(virtual reality)、ホログラフ

5.備　考

将来、チケット予約は、グローバルな情報サービスシステムの１つの機能として提供されるようになるだろう。その頃には、大規模な情報アクセスのためのインフラストラクチャが構築され、その上でさまざまなコミュニケーションチャネルがサポートされているだろう。今後、コミュニケーションはマルチモーダル化し、音声と他のモダリティとのインタラクションを考慮する必要があるだろう。特に何をどのようなモダリティで表現するかを適切に決める技術が必要となる。逆説的ではあるが、音声単独としての相対的なの重要性は低下すると考えられる。

参考文献

1.[Bobrow 77] D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H.Thompson, and T. Winograd. GUS, a frame-driven dialog system. Artificial Intelligence, 8(2):155-173, 1977.

2.[Lenat 86] D. Lenat, M. Prakash, and M. Shepherd. CYC: Using commosense knowledge to over come brittleness and knowledge acquisitionbottlenecks. AI Magazine, 6(4):65-85, 1986.

3.[Winograd 72] T. Winograd. Understanding Natural Language. Academic Press, 1972. Ph.D. Thesis.

4.[伊藤 92] 伊藤克亘. 音声対話システム. 電子情報通信学会技術報告、SP92-38:23-30、1992.

5.[長尾 93] 長尾確and 竹内彰一. コンピュータとの自然な対話のための新しいモダリティ-表情つきっk音声対話システムの試作と実験-. 人工知能学会研究会資料、 SIG-SLUD-9204:9-16, 1993.

3.6　計算機用音声インタフェース

3.6.1　はじめに

近年、計算機の分野では、従来の大型汎用計算機にかわり、それと同等以上の性能を持つ個人向けのワークステーションあるいはパーソナルコンピュータが全盛期を迎えようとしている。

将来的にもこの傾向は続くものと予想され、1000ＭＩＰＳの性能をもつデスクトップワークステーションが登場するのも時間の問題であると言われている。

従来の大型汎用計算機では、バッチ処理あるいは遠隔端末機からの定型オンライン処理が主であった。しかし、今日のワークステーション等では、従来の使用形態に加え、特に、個人の知的な生産活動を支援するための高度なＨＣＩ（Human Computer Interface あるいは Human Computer Interaction）が求められている。現在のワークステーション等では、キーボード、ディスプレイ装置、およびマウスを用いて、各々のデバイスの特性を生かしたＨＣＩが実現されている。ビットマップディスプレイとマウスを用いたＧＵＩ（Graphical User Interface）［Shuneiderman 87］は、従来から使われてきたキーボードによるコマンド入力に比較し、格段に使いやすく効率的なＨＣＩであり、今やほとんど全てのワークステーションやパーソナルコンピュータに採用されている。

現在のＧＵＩでは、手を使う入力手段として、言語情報を入力するキーボードと、位置情報を入力するマウスとが用意されている。この環境では、言語情報を入力している途中で、位置情報を入力しようとしすると、片手をマウスに持ち替えなければならず、その時点で思考の断裂が発生するという問題点があり、必ずしも人間の知的な生産活動をサポートするという究極の目標を実現しているとは言えない。

この点を解決しようとしたペン入力コンピュータの研究も盛んに行われている［中山 91］。ペン入力コンピュータでは、言語情報と位置情報が一つのペンから入力できるため思考の断裂がない。また、ペン入力コンピュータは、文字認識技術を用いることによりキーボードを不要にできるため、コンピュータを小型化できる点、あるいはキーボードに不慣れなユーザにも使える点にメリットがあると言われている。しかし、文字認識技術を使ってペンから言語情報を入力する速度はキーボードに比較すると極めて遅く、人間の思考を鈍らせてしまう。言語情報の入力には、ペン入力よりもキーボードの方が快適である。

より効率的な言語情報入力手段として有望なのは、音声認識である。音声は、熟練者でなくても、言語情報の入力速度がキーボードより早く、ペンやマウスなど位置指定デバイスとの相性も悪くない。音声認識は人間側の思考の断裂が少ないＨＣＩであると言える。また、画面表示を使わずに言語情報を伝える方法として、音声合成は唯一のものであると思われる。しかし、さまざまな問題点により、現在音声認識や音声合成を用いた計算機用音声インタフェースがその地位を確立しているとは言い難い。

以下、理想的な計算機用音声インタフェースの実現に向けて、特に個人ユーザが使う観点から、利用形態の例、必要な機能、装置形態、および研究課題について述べる。

3.4.2 利用形態の例

(1) マルチウィンドウシステムでの利用

マルチウィンドウシステムでは、ユーザは、複数のアプリケーションを同時に起動し、アプリケーション間を渡り歩くことにより、知的生産活動を効率良く進めていく。現状では、アプリケーションを切り替えるために、マウスでアプリケーションウィンドウをクリックしている。アプリケーションがキーボードを利用している場合、一次的に片手がキーボードから離れることになり、作業の連続性が極めて良くない。ユーザがアプリケーションの名前を発声し、システムがそれを認識することによりウィンドウの切り替えを行えれば、ユーザは両手をキーボードに置いたまま作業を継続することができ、より効率的な作業環境が実現される。

(2) プログラム開発時のコメント入力

プログラム開発時には、後々のメンテナンスのためにコメント行を挿入することが通常よく行われる。コメントの善し悪しがプログラムのメンテナンスの容易さを左右する。詳細なコメントはプログラムの局所的な説明として必要であるが、プログラム全体を把握するためには邪魔になることがある。

そこで、最近では、ハイパーテキスト構造がよく利用される。これは、プログラム本体のコメント行に、その部分の詳細なコメントの入ったファイルや関連情報の入ったファイルへのポインタを格納するもので、専用のエディタあるいはビューアを用いれば、コメントを編集／表示したい時だけ、その部分に関する詳細なコメントやそれに関連する仕様書の情報を見ることができる。

ところが、プログラムのコーディング中に各部の詳細なコメントをキーボードから入力することにすると、プログラムのコーディングに集中できなくなる。また、後から詳細なコメントを入れることにすると、コーディング時に考えていた内容を思い出せなくなる可能性がある。

そこで、コメントを音声で入力することが考えられる。音声の情報速度はキーボードのそれに比べ格段に早く、音声入力はコーディング作業を著しく中断する要因とはならない。この場合、コメント音声を認識し、テキストのコメントが自動的に生成できるにこしたことはない。しかし、もし認識誤りの修正作業が必要になると、コーディング作業を中断しないという本来の主旨に合わないため、かならずしも音声認識を使う必要はない。コメント行に置くポインタにコメント音声を録音したファイルを指定し（ハイパーメディア構造）、後でそれを聞いてテキストに変換すればよい。

ここでは、プログラム開発時に限って説明したが、コメントやメモを残す必要のある作業には、上と同じ方法が適用できる。

（3）　メニュー選択

現在のＧＵＩでは、マウスとプルダウンメニューによるアプリケーションの機能選択が良く用いられる。例えば、ワープロで文字の下に線を付けたい場合、キーボードから片手をマウスに移し、プルダウンメニューを表示させ、さらにマウスで「下線」の機能を選択し、最後にまた手をキーボードに戻すという一連の操作が行われる。これでは、思考の断劣が甚だしい。

機能選択を音声認識で行うことにより、キーボードから片手を離すことなしに作業を続けられ、思考の断劣が極めて少なくなることは明らかである。ここで注意が必要な点は、プルダウンメニューは、機能を選択するだけでなく、全ての機能を表示して見せるという一種のヘルプ機能をもっているという点である。この点を無視して単に音声に置き換えても、必ずしも使いやすくならない。

ＣＡＤ作業における音声によるメニュー選択の効果がShutohら［Shutoh 84］により報告されている。

（4）電子ニュース・電子メールの読み上げ

文字表示装置を使わずに、電子ニュースや電子メールの内容を知るには、音声合成を用いるかしかない。表示画面に見ずに他の作業をしながらこれらの内容をラジオ放送の様に聞き流すことができれば、仕事の能率を上げることができる。パソコン通信センターにおけ電子メールの読みあげサービスは、既に実用化されている［Kaseda 92］。

3.6.3　必要な機能

（1）１にかっこよさ、２に楽しさ、そして機能

従来は音声認識機能第一主義ですべてが考えられてきたが、個人ユーザにそれをあてはめても成功しないと思われる。まず、音声インタフェースを使っているユーザを見て、私も使ってみたいと思わせる「かっこよさ」が必要である。どんなに機能が高くてもかっこの悪いものは、個人ユーザには受け入れられない。恋人の前で醜態をさらさせる様なことがあれば、２度と使ってもらえないのは自明である。また、使っていて楽しくないといけない。むしゃくしゃする時にでも、音声インタフェースを使うと気持ちが晴れるくらいの「楽しさ」が必要である。最後に、当然のことながら従来と同様に機能の低いものは継続的に使ってもらえない。

（2）ハンズフリーマイク

ヘッドセットがかっこいいと思うユーザもいるかもしれないが、やはりうっとうしいものである。コマンドを１つ入力する場合にも、わざわざヘッドセットを付けないといけないようではユーザに使ってもらえない。

（３）耐雑音

ワークステーション等が置かれている場所は、事務室である。事務室には、空調機、電話のベル、人の話し声、構内放送などの音が満ちている。音声インタフェースに要求される機能は、これらの不要な音を取り除き、認識すべき音声だけを取り出して認識する機能である。

（４）認識開始指示

従来は、認識開始前に「認識スイッチ」をオンにしていた。これでは、使いやすい音声インタフェースは実現できない。

「認識スイッチ」のかわりに、常時計算機側でキーワード検出状態（一単語検出状態）になっており、キーワード発声を検出すると通常の認識状態に入る機能が、すでに自動車内の音声認識装置やＡｐｐｌｅ社の実験システム「キャスパー」に搭載されている。しかし、これによると当然のことながら発声量が増えること、ユーザに音声認識の存在を強く意識させてしまうなど問題が多い。理想的には、入力音声だけでなくユーザの全ての状況を音声認識側が常に把握し、アプリケーション側で認識対象となっている単語のみを的確検出し認識する機能が必要である。

（５）クイックスタート

従来のキーボード等の入力装置は、それをワークステーションに接続すると、すぐに使うことができる。個人ユーザは、箱を開けパワーオンしてすぐ使えることが当然だと考えている。音声インタフェースにも同様の機能が要求される。すなわち、使用前に音声登録を必要とする従来の特定話者認識方式あるいは話者適応方式では、いかに性能が高くても、個人ユーザに広く受け入れられるものは実現できない。また、立ち上げ時間が早い必要もあり、音声テンプレートあるいは音声モデルは装置内部にＲＯＭ化されているか、瞬時に記憶媒体からダウンロードできる必要がある。

（６）柔軟な認識誤り訂正機能

どんなにパターン認識の技術が進歩しても、認識誤りを無くすことは困難である。高い認識率の認識方式を採用するのは当然のことであるが、アプリケーションの作りが認識率100％を想定したものでは、使いものにならない。さらに、人間側の発声誤り率の方が、認識装置の誤

り率より高いこともありうる。従来のキーボードの文字認識率は限り無く100％に近いが、それでもバックスペースキーを備えている。バックスペースキーのないキーボードは、誰も使わないことは容易に想像できる。音声インタフェースにおても、柔軟な誤り訂正機能、あるいはアプリケーションでの誤り対処が必要である。これらの機能が充実していれば、95％程度の単語認識率で実用的なシステムが実現できると思われる。

（7）任意語彙認識

当然のことながら、アプリケーションごとに、音声認識対象語彙は異なる。そのため、任意の語彙を認識できる機能が必要がある。語彙セットは、アプリケーションごとに設定することになる。マルチウィンドウ型のワークステーションでは、現在どのアプリケーションがアクティブになっているかを知る機能も必要である。語彙セットの大きさは、アプリケーションによるが、1000単語程度が必要である。同時に認識できる語彙数は100～200語で良いことが多い。しかし、地名や人名を入力する場合は同時に数万語が認識できる必要があり、大語彙音声認識あるいは超大語彙音声認識の機能が必要である。

（8）連続音声認識

連続音声認識に関しては、電話番号やクレジットカード番号等の入力を必要とするアプリケーションが多いと考えられるため、連続数字認識機能は必須である。しかし、文章の連続音声認識はできるにこしたことはないが、それができるできないによって、音声インタフェースの価値が大きく変わるものではない。

（9）自然発生的音声の認識

連続音声認識よりも重要な問題が、自然発生的な音声を認識する問題である。例えば、「東京都（トーキョート）」という地名単語を入力する場合、「トーキョート」と文字どおりに発声すれば従来の認識方式で十分である。しかし、実際には「トーキョー」と発声する場合もあれば、「エート、トーキョー」と発声する場合もある。これらを如何に「東京都」という認識結果に結び付けるかが実用上非常に重要である。個別対処的に処理することは容易であるが、一般的な処理法の実現には極めて困難な問題が含まれており、今後の研究開発に期待するところが大きい。

（１０）自然な音声合成音

テキストを音声に変換する規則音声合成が実用化されて久しい。合成音の明瞭度（合成音を人間が聞いた時の認識率）は、継続的な研究の末かなり高いものが得られるようになってきて

いる。しかし、依然として合成音の自然性が高くないため、特定の用途にしか使われていない。理想的な音声インタフェースを実現するためには、この自然性を飛躍的に向上させる必要がある。

また、事務所内で音声合成を使う場合、あちこちで合成音がしたのでは、やかましくてしかたがない。ユーザひとりにしか聞こえないような機能が必要である。

3.6.4 装置形態

従来の音声認識／合成の形態は、計算機本体の処理能力が低かったため、専用の「音声認識／合成装置」あるいは「音声認識／合成ボード」という形態であった。しかし、近年、ワークステーション等の性能が飛躍的に向上し、また、マルチメディア・ワークステーションと称しＡＤ／ＤＡ変換器さらにはＤＳＰ（Digital Signal Processor）を内蔵したものが現れている。今後もこの傾向は続くものと思われ、1000ＭＩＰＳの性能をもつデスクトップワークステーションが登場するのも時間の問題である。将来は、音声認識は現在の仮名漢字変換ＦＥＰ（Front End Processor）と同レベルの音声認識ＦＥＰという形態になるであろう。音声合成は従来のプリンタドライバと同レベルの音声合成ドライバ（スピーカドライバ）になるであろう。すなわち認識／合成装置というハードウェアは姿を消し、ソフトウェアの形になると考えられる。

音声認識／合成のソフト化により、製品の提供側にとっては、膨大な開発工数が必要なハードウェアがなくなることで、製品開発コストを低く抑えることができる。ユーザ側にとっては、安価に音声認識／合成が手に入るだけでなく、最新の研究開発成果をいち早く利用できるというメリットもある。

連続音声認識や概念からの音声合成など、膨大な処理量や大きな知識ベースが必要とされる音声処理については、各ワークステーションでの処理ができないため、ネットワーク環境下で利用できる「音声認識／合成サーバ」という形態になると考えられる。この場合も、各ワークステーション側の音声認識ＦＥＰ／音声合成ドライバがクライアントとなり、音声認識／合成サーバを利用するという形態が考えられる。

3.6.5 研究課題

ここでは、上で述べた機能および装置形態を実現するために必要な研究課題について述べる。

まず個人ユーザを対象とする場合、音声インタフェースの「かっこよさ」と「楽しさ」を探る研究が必要である。これについては、現在までほとんど研究がなされておらず、あらたな研究が必要である。また、従来の音声認識装置購買者が認識装置をなぜ継続的に使わないかの理由の内、「認識率が低い」「雑音に弱い」以外の理由にも、一度目を向ける必要があろう。

効率追求の立場からは、音声インタフェースとキーボード、マウスなど他の入力手段との相性の研究が必要である。思考の断裂の少ないマルチモーダルなＨＣＩを実現するための研究である。これに関しては、例えば［Cohen 91］など盛んに研究されている。

ハンズフリーマイクと併せて耐雑音機能の研究は重要である。従来は、音声そのものの認識に研究勢力のほとんどを割いてきた。しかし、マイクに入力される信号の内訳は、認識対象となる音声は極一部で、それ以外のものがほとんどである。計算機用音声インタフェースを実現するためには、入力される信号の割合に応じた研究勢力の配分が必要であろう。

従来に比較して格段に高性能な不特定話者音声認識の研究開発は不可決である。不特定話者音声認識の実現のためには、

(1) 数千～数万人規模の多数話者音声データベースの構築と、
(2) 莫大なデータに含まれる多数話者音声情報をできるだけ欠損なく圧縮する技術と
認識時にそれを高速に取り出し入力音声と照合する技術

が必要である。(1)については、一研究機関では実施できないため、国家的施策が必要である。また、データベースの利用契約についても、従来のような安価な研究用データベス提供としての契約だけでなくで、一定の対価を払えばデータを商用に使用できる契約形態も必要となるであろう。(2)については、従来のマルチテンプレート法、HMM法、ニューラルネットワークを用いた方法などを基礎に、将来予想される計算機の性能向上や大規模音声データベースを十分に活かすことができる次世代の認識方式の研究開発が望まれる。

柔軟な認識誤り訂正方式の開発が必要である。これには、音声認識側で対処する方法とアプリケーション側で対話制御技術を用いて対処する方法とがある。一般的には後者の方が各アプリケーションに適した誤り訂正が行える可能性がある。しかし、誤り訂正手段がアプリケーションごとに異なるとユーザが混乱する恐れもあり、誤り訂正方式のガイドラインを制定する必要がある。また、すべてのアプリケーションが音声認識を想定していないことも考えられるため、アプリケーションによらない音声認識側での誤り訂正方式の研究も必要である。これに関しては、音声認識ＦＥＰとアプリケーションの間に存在して、アプリケーション側でできることと音声認識結果を総合判断し、人間と対話を行いながら認識結果を修正していく「エージェント」の考え方が有望である。

音声合成音の自然性改善については、従来のフォルマント合成やＰＡＲＣＯＲ合成等の音声のスペクトルモデルや、従来のピッチ周波数の時間変化モデルでは、明瞭性は達成できても、自然性は限界に達していると考えられる。より高度な音声モデルの研究開発、あるいは音声モデルをできるだけ使わない波形編集方式などの蓄積再生方式的あるいはサンプリング音源的な

発想に基づく音声合成方式の研究開発が必要である。

最後に、音声認識／合成装置の形態としては、将来はソフトウェアという形態になると考えられる。ここでの重要な課題は、音声認識ＦＥＰ、音声合成ドライバ、オペレーティングシステム、およびアプリケーションプログラム、それぞれの間のインタフェースの標準化である。既存のＡＰＩ（Application Programming Interface）を流用するのではなく、音声インタフェースの特質を活かせるＡＰＩさらにはＡＢＩ（Application Binary Interface）の標準化が研究課題となる。この標準化により、Ａ社のアプリケーションをＢ社の音声認識ＦＥＰで動かすことなどが可能になる。ユーザは価格優先あるいは性能優先などの各自の目的に応じてシステム構築を行うことができ、ひいては音声インタフェースが広く世の中で使われることにつながると期待できる。

また、音声認識／合成サーバに関しては、サービスプロトコルの標準化の問題は音声認識ＦＥＰ等と同様である。さらに、サーバにはワークステーションやパーソナルコンピュータなどのさまざまな性能のクライアントとの接続が予想される。そのため、各ユーザに均質のサービスを提供するためには、各クライアントの性能に応じて認識処理をクライアント側とサーバ側に動的に切り分ける技術の開発が必要となろう。

３.６.６　おわりに

ここでは、計算機用音声インタフェースについて述べた。繰り返しになるかもしれないが、今後、計算機のダウンサイジングおよび高性能化という傾向が続くことは間違いなく、そこで使われるＨＣＩとして音声認識／合成は極めて有望である。そして、音声認識／合成を現在のＧＵＩ並みのＨＣＩに育てあげるためには、従来の音声認識研究の範疇で捉えられている技術は勿論のこと、広くコンピュータ技術全般、さらには情報家電技術との係わりを研究していく必要がある。

参考文献

1. [Shuneiderman 87] Shuneiderman, Ben, (東、井関監訳) "ユーザインターフェース設計"、日経ＢＰ社、1987
2. [Shutoh 84] Shutoh, Tomoki, et al., "Voice Operation in CAD System", In Hendrick, H. W., and Brown, O., Jr., (Editors), Human Factors in Organization Design and Management, Elsevier Science Publishers B. V.(North-Holland), Amsterdam (1984), pp.645-649
3. [Cohen 91] Cohen, R., Philip, "The Role of Natural Language in a Multimodal

Interface", 91 International Symposium on Next Generation Human Interface, November, 1991, Tokyo.

4. [中山 92] 中山俊一、"ペン入力コンピュータが、日本市場開拓に踏み出す"、日経エレクトロニクス、no. 557、6月22日号, 1992, pp.135-142

5. [Kaseda 92] Kaseda, Mitsuko, et al., "Speech Service for Electronic Mail Using Speech Synthesis Technology", Voice Systems Worldwide '92 Japan, July, 1992, pp.69-73

3.7 携帯端末

3.7.1 背 景

「いつでも、どこでも使える」をキャッチフレーズに携帯型の情報通信端末の普及がはじまろうとしている。その代表は携帯電話、自動車電話でありアナログ方式はすでにサービスが開始されており、ディジタル方式も93年からサービス開始の予定である。その市場規模は2000年には2000万台を超えるともいわれており、情報家電メーカからの期待も大きい。携帯電話、自動車電話では提供される機能は電話機能だけである。一方、情報分野ではポストノートパソコン、あるいは電子手帳上位機種として携帯型の情報端末への期待が高まりつつある。93年にはApple社が携帯端末「Newton」を発売するのをはじめとして、いくつかのメーカが携帯端末の発売を計画している。

携帯端末では手のひらに収まる大きさという制限からキーボードを設置することができず、ペン入力への期待が高まり、すでにペン入力OS等も開発され、上記「Newton」等にも搭載されている。さらに進んで、音声を入出力媒体として使うことを考えると、ペンやディスプレイすら必要なくなる。

携帯端末は電子手帳のようにスタンドアロンで利用する場合、ネットワークを介してデータベースにアクセスして利用する場合がある。後者では無線ネットワークの整備が必須の要件となるが現状の携帯電話、自動車電話などの無線ネットワークでは送信のために必要な電力が大きいという問題がある。現在、無線ネットワークとして期待されているのは次世代コードレスホンであるPHP（パーソナルハンディホン）である。消費電力が0.01Wと小さく、端末を小型軽量化できるとの期待が大きい。データベースにアクセスしてサービスを受けるようなことを考えるとデータベースサービスの整備も必須となる。現在、データベースサービスの整備はまだ不十分であり、また、PHPのサービスがはじまるのは早くても94年以降といわれており条件が整っている段階ではない。しかし、これらは今後の必然の方向と考えられ、5年先、10年先にはかなり環境が整っているものと考えられる。

ここでは、こうした携帯端末に音声入出力機能を持たせたときの使用形態、必要機能、技術課題等について考察する。

3.7.2 携帯端末における音声入出力

まず各種応用を携帯端末に搭載した場合を想定し、使用形態（誰がどこでいつ使うのか）、携帯の必要性などを考察する。応用としては本章の他の節でとりあげている次の7つの応用システム、すなわち（1）文書作成、（2）ロボット、（3）CAI、（4）情報案内、（5）予約、（6）インターフェース、（7）ナビゲーションシステム、および携帯端末特有の応用として考えられる（8）携帯電話（9）電子秘書（電子手帳上位機能）について考える。検討

結果を表3.7-1に示す。

表3.7-1　携帯端末の使用形態

	応　用	使用形態		携帯の必要性	備　　考
		使用者層	場所／時		
1	文書作成	ビジネスマン	出先／移動中	○	出先でのメモ作成
2	ロボット	主婦	家庭	△	リモコン（家庭内で携帯）
3	CAI	先生／学生	学校，家庭	△	通学途中に使う？
4	情報案内	*	*	△	
5	予約	*	*	△	
6	インターフェース	*	*	△	
7	ナビゲーション	運転者	車内／運転中	◎	車載（車から取り外して使う？）
8	携帯電話	ビジネスマン	出先／*	◎	音声ダイアル，接続後上記2～6の処理
9	電子秘書	ビジネスマン	出先／*	◎	スケジュール管理等

表中の*は、特に使用者層、場所、時間等を特定できないことを示す。現在の携帯電話の使用者層からも予想されるように、主な使用者層はビジネスマンであると考えられるが、たとえば学生が通学途中に試験の予習をするといった使い方も考えられる。また、個人ごとに携帯端末を持ち、その中に家庭内電化機器のリモコン機能が統合され、音声コマンドで制御できるといった形も考えられる。

3.7.3　必要機能

携帯端末は応用システムというよりは使用形態という側面を持っている。すなわち携帯端末の上にどの様な応用を載せるか次第で携帯端末に持たせるべき機能は異なってくる。各種応用ごとに必要な音声機能の詳細については本章の他の節にゆずり、ここでは携帯端末において必要な音声機能について簡単に触れ、また携帯端末に固有な音声機能以外の問題について述べる。

（1）音声認識機能

携帯端末においても出先での簡単なメモ作成など必要になろう。ただし、ワープロに要求されるような高機能は必要ないと思われる。たとえば、仮名文字だけでもメモとしては十分に役に立つ。この場合仮名漢字変換は不要となり、辞書を搭載する必要もなくなる。音声ダイアルや簡単な問い合わせであれば小語彙の単語認識でも十分であろうが、やはり、使いやすさを考えると大語彙化、話し言葉の理解が期待される。これは、端末に搭載可能な計算パワーやメモリ量とのトレードオフとなろう。

（2）音声出力機能

ディスプレイのない場合を考えると情報出力手段は音声出力しかなく、基本的に任意の文章の音声合成が必要となる。

（3）端末機能

携帯端末の端末単体として要求される機能について簡単に触れる。

小型軽量　：　まず、携帯性のために小型軽量ということが要求される。小型というのは手のひらに収まるサイズあるいはポケットサイズ。また、重量の目安としては500 g 以下がめどになる。

低消費電力：　携帯端末はバッテリーによって駆動される。連続的にある時間使うことを考えると低消費電力が必須の要求となる。連続8時間使用を想定すると要求される消費電力は0.01Wオーダーとなる。

無線通信　：　ネットワークを介して使う場合に無線通信機能は必須である。小型軽量低消費電力を満たしつつ、無線通信機能を搭載する必要がある。

（4）環境（インフラ）整備

電子手帳のようにスタンドアロンで使う場合は別として、ネットワークを介して使う場合にインフラの整備は必須である。1つは、無線ネットワークの整備であり、さらにそれを介してサービスを受けるために各種サービスの充実が要求される。

3.7.4　装置形態

携帯端末に音声入出力機能を持たせるとき、装置形態としては端末側に音声入出力機能を持たせる場合とネットワーク上のセンターシステム側に持たせる場合の2つに大きく分けて考えることができる。どちらの形態をとるかは端末側とセンター側の負荷分担の考え方や制度上の問題、プロセッサのコスト等によって判断すべきものであり、どちらが良いとは一概にいえない。ここではそれぞれの形態について述べる。

（1）端末側

端末側に音声機能を持たせる場合、音声認識のために必要な音響モデル、辞書、言語モデル、アプリケーションに必要な個人データベースなどを端末内に持つ必要がある。このためには数

十ＭＢのメモリを端末内に持つ必要が生じてくる。現状端末内に持つことのできるメモリサイズはせいぜい数百ＫＢである。半導体の進歩に待つところが大きいがすでに256ＭＤＲＡＭが発表されており、５～10年先を考えると数十ＭＢを端末内に持つことも夢ではない。ただし、データ量削減の努力は常に必要であり、音響モデルのデータ圧縮表現技術、データ共用化技術等が必要になる。また、メモリ量の観点ばかりでなく、処理の高速化も要求される。最低数十ＭＩＰＳ、高機能化を求めれば数百ＭＩＰＳを越える高速プロセッサを端末内に持つ必要がある。もちろん、各種処理量削減策を施していくことが必須である。予備選択法、ビームサーチ等処理量削減技術の導入や改善、統合が必要である。

一方、端末側に音声認識機能を持たせる場合には特に不特定話者音声認識機能は要求されず、特定話者で十分である。むしろ、話者認識機能と音声認識機能を融合し、特定個人しか使えないようにする技術の開発も必要になろう。

（２）センター側

センター側に音声機能を持たせる場合、ネットワークを介して音声を飛ばし、センター側で音声認識をおこない、センター側で合成した音声をネットワークを介して端末で受け取ることとなる。ＰＨＰの場合、各親機に音声認識、合成機能を持たせることになろう。

センターあるいは親機に認識機能を持たせる場合、誰でもが使える不特定話者にするか、ユーザＩＤごとに標準パターンを登録できるようにシステム的に対応するかの問題が生じてくる。また、無線環境に依存する劣化音声の認識の問題が生じてくる。

３.７.５　技術課題

携帯端末にどのような応用を載せるかにより携帯端末の音声入出力における技術課題も異なってくる。各種応用に固有の課題については本章の他の節で詳細に述べられているので、ここでは携帯端末に特有な課題を中心に述べる。

（１）大語彙

携帯端末でも、計算パワーやメモリ量の許される限り大語彙化が望まれる。大語彙化の問題は本章の他の節で詳しく述べられているのでここではこれ以上述べないが、携帯端末では計算パワーやメモリ量が制限され、処理量／メモリ量削減技術が特に重要になることを指摘しておく。

（２）雑音対策

携帯端末はどのような場面で使用されるか特定できず、様々な雑音環境で用いられることを

想定しなければならない。車内騒音、駅構内の騒音、歩行中の騒音など様々な騒音下で安定に動作する必要がある。このような、様々に変わる環境に適応するための技術が必要となる。適応フィルタを適用するのも1つの方向であろうが、想定される使用環境の騒音辞書を用意しておくといった方法も考えられる。

携帯端末では雑音による音声パターンの変形だけでなく、音声区間検出の問題が厳しい問題となる。たとえば、歩きながら携帯端末を使う場合を考えると、マイクと口との位置関係は一定に固定されず変動する。これにしたがって音声パワーにも変動が生じてしまう。こうした変動を吸収して音声区間検出をおこなう技術の開発が必要である。

（3）ささやき声

携帯端末は様々な場面で使用される。公共の場で用いるとき、周囲の人に聞かれるのがはばかれるような場合では必然的に小さな声で話す。極端な場合にはささやき声を認識する必要がある。雑音下ではユーザの発声が変形するというロンバード効果の問題があるがささやき声はさらに極端な発声変形と考えられる。発声の全体にわたって無声化現象が生じるものと考えられるが、スペクトルの変形の状況や、音響モデル作成の方法、音声検出手法など研究する必要がある。

（4）話し言葉、対話処理

携帯端末の理想形態はApple社のコンセプトであるＰＤＡ（Personal Digital Assistance）にもあるように個人向け電子秘書であろう。簡単ではあるが面倒な仕事は全てエージェント（代行者）に任せ、自分は高度な仕事に専念する。スケジュールの管理、電話の発呼、手紙代筆等を言葉で指示しておこなわせる。言葉で指示しておこなわせる（代行させる）ということを前提とすると、話し言葉理解あるいは自発的発話の理解が重要な課題となる。また、こうした話し言葉でのコミュニケーションでは誤解や曖昧性がつきものである。こうした状況で確実にコミュニケーションを図るには対話が欠かせない。対話管理技術も携帯端末にとって重要な技術の１つである。

3.7.6　まとめ

携帯端末に音声入出力機能を持たせる場合の使用形態、必要機能、技術課題について考察した。最後に5年後、10年後の要求レベルを現状レベルも含めて表3.7-2にまとめて示す。

表３．７-２　携帯端末の技術課題

重要研究課題		技術水準		
		現　在	５年後	１０年後
音声	大語彙 雑音対策 ささやき声 話し言葉，対話管理	～100語，～10文 ～１０ｄＢ，定常 × 研究着手	～500語，～50文 ～０ｄＢ，定常 ? ?	～3000語，～300文 ? ? ?
他分野	ＣＰＵ（MIPS） メモリ（Byte） インフラ	１０ ２００ｋ セルラ	１００ ２Ｍ ＰＨＰ	５００ ２０Ｍ ?

3.8　マルチメディア（音声）データベース検索システム

データベース（DB）のマルチメディア化が進められている。音声DBの検索は他のメディアと同様、当面はヘッダー情報を用いておこなわざるをえない。しかしマルチメディアDBの増大に伴い、ユーザーに負担がかかる現行のDB利用システムから、音声による直接検索のできるシステムへと移行することが望まれる。

ここでは音声DB検索システムの概要、機能、研究課題についてまとめる。

3.8.1　音声DB検索システムの概要　―　構成と利点　―

図3.8-1にマルチメディアDB（MMDB）検索システムの構成を示す。システムは、マルチメディアDB（光ディスクファイルシステム）、マルチメディアDB管理サーバ(MM-DBMS)、および複数のマルチメディアDB利用端末からなる。

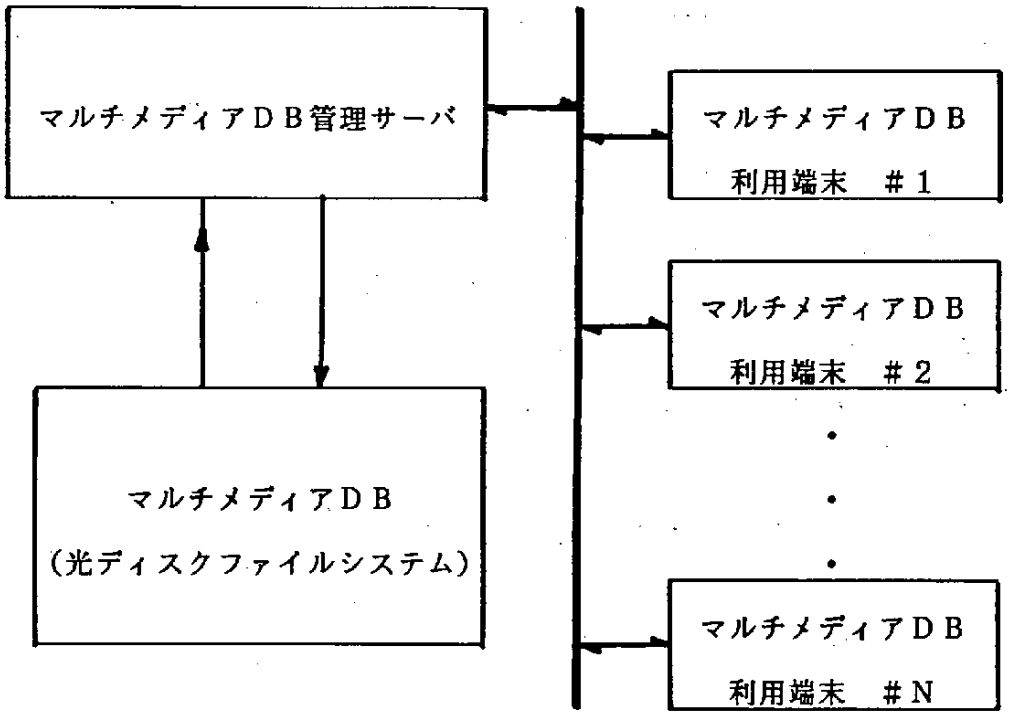


図3.8-1　マルチメディアDB検索システムの構成

このようなシステムに、将来、音声ファイルを直接利用者の音声で検索する機能が加わると、現行のキーワードによる検索システムと比較して、次のような利点を持つと考えられる。

（イ）　音声DBに直接アクセスできる

（ロ）　キーワード体系（シソーラス）を知らなくても利用できる

（ハ）　DB管理システムに依存しない

（ニ）　ユーザーインタフェース（UI）に優れている

このようなマルチメディアDB検索システムは、大規模な専用システムとしてのほか、汎用のワークステーション上に実現されるようになるであろう。

3.8.2　音声DB検索システムの機能

音声ＤＢ検索システムに必要とされる機能には、

（1）　キーワード自動抽出

（2）　ワードスポッティング機能

（3）　音声データの自動変換機能

（4）　異話者間の音声適合機能

（5）　特定音声復元機能

（6）　使いやすいＵＩ機能

がある。

（1）は連続音声DBから、自動的にキーワードを抽出する機能である。これによって、音声DBをあらかじめ構造化し、迅速な検索がおこなえるようになる。（2）は決められたキーワードを検索（スポッティング）する機能である。利用者の指定する（発声した）キーワードの検索に使われるほか、音声DB自身のキーワード自動抽出にも必要な機能である。

利用者の音声と、DB中の音声との差異には次のようなものがある。

A）　音声データの仕様の違い（サンプリング周波数／ビット幅／圧縮方法）

B）　話者の違い（言語／男女／年齢／方言…）

C）　収音環境の違い（マイク／騒音／…）

音声DB検索では、これらの差異を克服することが要求される。音声DBのデータフォーマットは、通常、異なる仕様で収録されている。このため、（3）の「音声データの自動変換機能」（異なる音声ファイル間の自動変換機能）が必要になる。また、（4）は上記B）の話者の違いを克服する機能、すなわち利用者の音声から音声DB中の話者への音声適合機能である。音声DBは、必ずしも音響的に良い環境で収録されたものばかりではない。雑音の他、楽音や複数の人の声が同時に収録されている場合もある。このような音声DBでは、（5）の特定音声復元機能が必要になる。

音声DB検索システムは、不特定多数の人が場合によっては遠隔地から利用することを前提としている。このため、（6）の使い勝手の良いUIを提供することが大切になる。

3.8.3 音声DB検索システムの研究課題

ここでは、音声DB検索システムの研究課題として、3.8.2で列記したシステムの6つの機能を実現するための方法を考察する。

（1） キーワードの自動抽出

テキストに対するキーワードの自動抽出は、現在、研究段階にある［池原 87］。このため、連続音声DBから自動的にキーワードを抽出することは、まだまだ先の技術になると考えられる。しかし、あらかじめキーワードを抽出してDBを構造化する機能は、DB管理システムの効率的な運営に不可欠である。当面は、テキストや画像、ことにテキストDBからのキーワード抽出を中心に、これらと連携した音声中のキーワードを検出することが研究課題となろう。

（2） ワードスポッティング［Wilcox 92］

この機能は、既知の単語（または文節・文）を音声DB中から抽出する際に必要となる。しかし、（1）に述べたキーワードの自動抽出では、未知語を抽出することが要求される。現在、研究は主として利用者と音声DB中の話者が同一人で、しかも既知の単語抽出を対象におこなわれている（同一話者のキーワードスポッティング）。しかし通常は、異話者の音声DBを利用することのほうが多いと考えられるため、異話者データによるキーワードスポッティングが重要な課題となる。

（3） 音声データの自動変換

音声DB検索に限らず、マルチメディアを分散環境で利用する場合、音声データとその属性（サンプリング周波数／ビット幅／圧縮方法など）は、一体として扱われる（オブジェクト化）。この場合、音声データの仕様の差異はオブジェクト変換を通して吸収される。マルチメディア・オブジェクト分散環境を実現するには、様々な規約を整備する必要があり、現在、この標準化作業がIMA（Interactive Multimedia Association）を中心に進められている。音声DB検索システムも、このフレームワークを利用することになろう。

（4） 異話者間音声適合

利用者音声から音声DB中の話者への適応は、音声認識における音声適合機能に相当する［小坂 92］。ただし、音声DB中の話者に対する適応は、原則として教師なしでおこなうことにな

る。このため、不特定話者音声に対するキーワード抽出が高い精度で実現されていることが前提になろう。

また、年齢／方言／男女／言語の違いなど多くの難問が控えており、一歩づつ適応能力を向上させる研究が大切である。

（5） 特定音声の復元（Speech Enhancement）

音声DBの収音環境は、様々であろうことが予想される。マイクの違いにはじまり、騒音環境の違い、さらには楽音の混合や複数の人の声が同時に収録されている場合もあろう。音響環境への適応、ノイズキャンセル、さらに人間の持つカクテルパーティー効果に相当する技術の開発など、多くの研究の積み重ねが必要とされる。

（6） UI環境

音声DB検索システムでは、不特定多数の人が遠隔地から、分散管理されたDBを利用する場合が多いと考えられる。今後整備されるオブジェクト分散環境下では、マルチメディアDBを利用する端末は、マルチメディアによるアクセスを可能にしなければならない。音声DBを直接、音声でアクセスすることは自然な形ではあるが、より自然な対話は音声・文字・画像といったマルチメディアを使ったアクセスであろう。マルチメディアを統合した対話環境、すなわちマルチモーダル対話環境の整備と研究は、まだはじまったばかりである［新田 92］。

当面の研究課題としては、対話におけるメディアの組み合わせの相性に関する研究、利用者とシステムとの対話におけるインタラクションとその深さの制御に関する研究などがあげられる。

参考文献

1. ［池原 87］ 池原、中園、白井：キーワード自動抽出システム、研究実用化報告、36、9. pp.1151-1158, 1987
2. ［Wilcox 92］ L. D. Wilcox and M. R. Bush: Training and Search Algorithms for an Interactive Wordspotting System, ICASSP'92, pp.II-97-100, 1992
3. ［小坂 92］小坂、鷹見、嵯峨山：話者混合SSSによる不特定話者認識と話者適応、信学技報SP92-52, PP.17-24, 1992.9
4. ［新田 92］ 新田：音声マルチモーダル対話の動向、人工知能学会研究会資料 SIG-SLUD-9202-2, 1992.7

3.9 ナビゲーションシステム

3.9.1 背 景

今までは、カーナビゲーションシステムは、必要性は感じられていても、実用にはほど遠いと考えられてきた。特に従来のものはジャイロやタイヤによる走行距離等から現在位置を割り出そうとしていたので誤差が大きく実用にならなかった。しかし、通信衛星の発達により、現在の自分の場所がかなりの精度で分かるようになってきた。さらにＶＩＣＳ（Vehicle Information and Communication System：道路交通情報通信システム）推進協議会が、建設省によるＲＡＣＳ（Road Automobile Communication System ：路車間情報システム）と郵政省によるＦＭ多重放送を利用するメディアサービスの両建てで新たな都市交通システムづくりに動き出すことになった。ナビゲーションに使われるＣＤ－ＲＯＭの規格も、自動車メーカが中心となってまとめたＣＤＣＲＡＦＴ（CD and CRT Applied Format）と、オーディオ機器メーカ等多数が参加するナビゲーションシステム研究会による統一フォーマットが検討されている。　このようにカーナビゲーションシステムは実用の時期を迎えようとしている。近年交通量は増え続けているが、装置が普及して的確な情報を提供できれば集中を避け、交通緩和にも役立つ。位置と進行方向がわかるので、知らない所でも道に迷わない。システムによっては行楽情報も検索できる。

カーナビゲーションを使うときは、運転者から視界や手足による操作を強制させることは危険である。この点、音声によるインタフェースはこの目的に最もかなったものである。ここでは、ナビゲーションシステムのインタフェースに音声を使うことを前提にして述べる。

3.9.2 ナビゲーションシステムの構成

音声によるカーナビゲーションシステムは以下のような構成になるだろう。

- 位置検知部
- 地図検索部
- 発話理解部
- 発話生成部
- 動作管理部

位置検知部はＧＰＳ（Global Positionning System）衛星による位置確認を行う。地図検索部は、ＣＤ－ＲＯＭ等に地図データが入っており、距離や道路名などの情報が入っている必要がある。発話理解部は音声認識を行う部分であるが、音声認識処理は常時行い、発話が入力されるたびに意図を解釈する。発話生成部は、発話理解部で基準以上のポーズが得られた場合に起動し、割り込みがあった時には停止する。あいづちが得られた場合は継続する。動作管理部

は、運転者との対話進行を管理する。

3.9.3 ナビゲーションにおける課題

位置検知部はＧＰＳ衛星による位置確認方法だと高いビルの谷間や衛星の少なくなる時間に受信ができないことがあるが、今後の課題であろう。地図検索部は、渋滞や事故、混雑度等の情報が日々変わるものであるので文字放送対応チューナ等で交通情報や天気予報が得られるものと併用することになると思われる。

発話生成部は、地名等を自由に発声できるようにするために規則合成型の合成器が便利である。ナビゲーションに必要な文型はかなり制限されたものになるので、ナビゲーション用にチューニングすれば、現在の技術でも利用できるだろう。

発話理解部では、ナビゲーション特有の問題が幾つかある。システムが発話している最中に割り込みをかけて、ユーザが発話してくる場合があるので、発話生成部の音声は打ち消して認識する機構が必要である。このような割り込みはユーザ主導のタスクで多く見られ、無視できない現象である。また割り込みのように利用者が積極的に対話しようとする場合とは逆に、何を話していいか分からず黙り込んでしまう場合も考えられる。この場合、システム側で利用者をうまく誘導するように対話を生成しなくてはならない。

さらに車内というノイズ環境にも留意されたものでなければならない。ノイズは、車の走行音だけでなく、オーディオ装置から出る音や同乗者の会話もある。同乗者の会話とシステムとの会話を区別するために、会話の内容を意味理解して判断することは非常に困難なことなので、音声の到来方向を調べたりする方が現実的である。

他にも、従来の研究では不要語として扱われてきた、あいづちが多く現れる。あいづちは、発話権の継続を許す重要な言葉である。認識を含めた応答速度もナビゲーションには重要な要素である。あまり応答が遅いと車が目的地を通り過ぎてしまうことになりかねない。発話のタイミングも重要な問題である。

ナビゲーションシステムは音声対話システムであるので、多くの対話システムと同様な問題が発生する。話し言葉に対応する文法や、言葉の省略などを吸収できるパーザが必要であるし、地名などの発話を許した場合、単語数が膨大になるので対話による絞り込みが必要になる。目的地に関する知識をあらかじめキーボード操作で入力するなどして、発話理解部の負担を減らすような工夫が有効であろう。

参考文献

1. ［菊池92］菊池英明、工藤育男、小林哲則、白井克彦、"ナビゲーションシステムにおける音声対話インターフェースの構成"、人工知能学会研究会（第３回）資料、pp.19-27

第４章　まとめ

第4章　まとめ

第3部では、音声認識·理解技術を応用システム側からの視点にたって必要となる要素技術、研究開発課題を考察してきた。ここでは、今後10年を見込んで有望と考えられる応用システムを取り上げた。個々のシステムについて、その使用形態や機能、またどのような問題があり、そのためどのような技術が必要となるかなどについて述べた。このような考察は、われわれが今後、音声認識·理解関連の研究開発プロジェクト等を計画立案していく際に、その将来像を具体的にイメージし、研究の方向や重点の置き場所を明確にするのに有用であると信ずるものである。

個々の応用システムには、それぞれ固有の問題があり、音声認識·理解技術の中でも優先されるべき要素技術が異なる場合、あるいはこれまでほとんど研究されてこなかった機能要素が要求される場合もある。すなわち、応用システム側からみると技術要素は単一の平面上にはなく、1つのシステムにインテグレートされる要素技術はその時点においては様々なレベルにある。また、表層においては単純そうにみえる機能も、それを実現するには極めて高度な技術が要求される場合がある。本報告書は、第一線の研究者によって書かれたものであり、これらの技術的な問題点が明確に踏まえられており、単なる将来イメージとしての応用システムの記述ではないことを協調しておきたい。

なお、冒頭でも述べたように、昨年度報告書で取り上げなかった要素技術の動向について、第2章で解説した。これによって、音声認識·理解関連技術をおおむね網羅したことになる。この数年の技術動向を知りたい読者の方々に広く役立てていただければ幸いである。

第４部　音声認識･理解技術の海外動向

第１章　米　国

第1章　米　国

1.1　概　要

米国を中心とする海外の音声認識･理解技術およびその関連技術の動向を調査するため、平成4年10月10日から2週間の日程で、カナダ、米国を訪問した。前半の週は音声言語処理に関する国際会議 International Conference on Spoken Language Processing（ICSLP）-92に参加し、世界の第一線の研究者の発表を聴講し、研究開発動向を調査した。後半の週は、昨年度の海外調査で訪問できなかった米国の先進的な研究機関であるCMU（Carnegie Mellon University）、ハーバード大学（Harvard University）、SRI（Stanford Research Institute）Internationalを訪問し、見学・意見交換をおこなった。本章では、はじめに米国の全体的な研究開発動向について述べ、後半で上記の3機関の訪問調査結果について報告する。

米国では、DARPAの下で音声言語理解プロジェクト、SLS（Spoken Language System）が強力に進められており、その予算規模は年間約1000万ドル余りといわれている。このプロジェクトにはMIT（マサチューセッツ工科大）、CMU（カーネギーメロン大）、SRI（スタンフォード研究所）、NIST（国立標準研究所）、BBN（ボルトベラクニューマン社）、TI（テキサスインスツルメント社）、AT&Tベル研究所などの音声研究グループが参加している。このプロジェクトでは、共通の音声データベースを構築し、さらに共通のタスクを設定して、各研究グループが競争的に研究をおこなっている。また、年1回のワークショップを開き研究発表と意見交換の場を設けている。このワークショップの論文集は、Proceedings of Speech and Natural Language Workshop（Morgan Kaufman Publishers）として市販されている。各研究グループのおこなっている研究内容については昨年度報告書におおむね記載されているので参照されたい。

SLSプロジェクトの音声理解システム用タスクとして、この数年、研究対象とされてきているのがATIS（Air Travel Information Service）である。ATISは、フライト情報に関する利用者からの問い合わせに答えるという設定のタスクであり、フライト情報に関するデータベースと音声理解システムが結合されている。ATISの語彙数は約1000語で、利用者は、いわゆる自由発話音声によって質問することができる。参加している研究グループは、その性能を競うことになるが、その性能の評価指標に関する研究もおこなわれている。データベースの管理や性能評価に関しては、国の機関であるNISTが統括している。

同じくSLSプロジェクトの下で大語彙ディクテーション用のタスクとして研究されているのがWSJ（Wall Street Journal）である。これは経済紙として知られる同紙の文章を読み上げたものである。5千語、2万語などのコーパスが用意されているが、語彙数は一般には、さらに多いので、語彙数に関して一種のスケーラブルなタスクとなっている。ドラゴンシステムズ社、

MIT、CMUなどがこの研究に参加している。

音声データコーパスとしては、他に、TI社によるSwitchboad Corpusとよぶ電話網を利用した会話音声の収集が進められている。また、Voice Across Americaとよぶ同じく電話網を使用した自動音声収集がある。これは、主に数字、名前など簡単な音声の大量人数データのコーパスである。

このSLSプロジェクトと同じ分野の研究についてみると、日本の研究機関の研究水準に比較して進んでいる点が多い。このプロジェクトに伴って研究されている音声理解システムの評価法なども日本ではまだ議論されておらず、参考にすべき点である。ただし、これらの米国の研究機関では、特徴抽出、音韻モデルなど基本的な部分が似た手法となってきており、しかも全体的なアプローチも似ており、手法上の独自性が結果にどの程度反映しているのか疑問が残る。

なお、同じくDARPAの下で言語と音声のデータベースに関連して、Linguistic Data Consortium（LDC）が組織されて、企業、大学、公的研究機関等へ上に述べたような音声データベースやテキストデータベースの提供がおこなわれている。こうしたデータベース等に関する組織的な動きについては、米国が明らかに先行している。

DARPAのプロジェクトに参加してない代表的な機関としては、IBMがある。同社は、70年代からHMMの手法にもとづいた音声認識の技術開発をおこなってきており、その技術水準は高く、大語彙ディクテーションシステムなどで成果をあげている。また、最近では、Apple社も音声認識・理解システムの開発に積極的である。

1.2　訪問研究機関

（1）カーネギーメロン大学（CMU）

CMUのSchool of Computer Science（Robotics Institute）を訪問した。音声の研究をおこなっているのは、Dept. of Electrical and Computer Eng.に所属するRichard Stern助教授のグループである。このグループは、DARPA（米国国防省、Advanced Research Project Agency）の音声言語システム（SLS）プロジェクトに参加している主要メンバーの1つである。

研究グループの中の中心メンバであるXuedong Huang博士から、SLSプロジェクトの音声理解システム用のタスクであるATISのデモ、ならびに現在進めている研究課題について紹介を受けた。ATISのタスクに用意されている語彙数は約1000語、平均分岐数60、単語についてのBigramが文法の1つとして与えられている。利用者の質問は、いわゆる自由発話で受け付けられる。ただし、割り込みキーを押し続けて発話中であることを示す。マイクロホンは卓上型またはヘッドセットの2タイプである。以上のATISに関する仕様はSLSのプロジェクトとして共通である。CMUのシステムは、数年前に開発されたSPHINXに改良を加えたもので、SPHINX-Ⅱと呼んでいる。この間、加えられた改良には、高次動的特徴（いわゆるデルタケ

プストラムの変化分）、話者正規化処理、Sub-phoneticな単位（Senone）の導入があり、特にSenoneの効果が大きいということである（図1.2-1参照）。

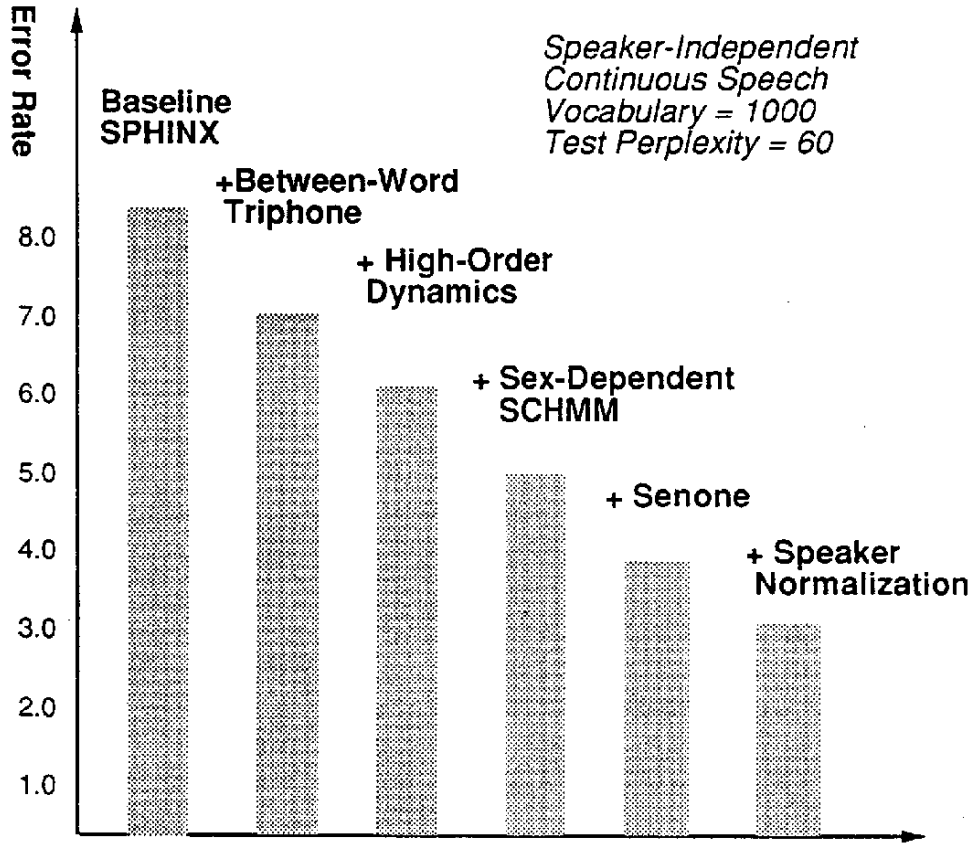


図1.2-1　CMUの音声認識システムSPHINX-Ⅱによる性能評価試験の結果。
各要素技術が認識性能に及ぼす影響を評価（X. Huang他によるレポート、"The Sphinx-II speech recognition system: an overview", CMU-CS-92-112より）。

システムの動作は、認識部分についてはほぼ実時間で、発話後数秒でフライト情報に関する応答が表示される。DARPAの中でのシステム性能評価については、特に音声認識の部分が高いという結果である。

今後の研究については、ATISは語彙数の点で不足であり、より語彙数の大きいWall Street Journalのタスクに研究対象を移していくということである。これに関連して、語彙に関して限定の無い認識システム、未知語処理などが今後の重点的な研究課題であると話していた。

CMUでは、Robotics Instituteの所長であるTakeo Kanade教授から同所の概要について、ま

た、Katsushi Ikeuchi博士から、関連テーマとしてコンピュータビジョン関係の研究（ヘリコプタの自動操縦、形状認識、組立て教示システムなど）について紹介を受けた。

（2）ハーバード大学

ハーバード大学では、Division of Applied Science（DAS）のRobotics Laboratoryを訪問した。DASは、大きく4つの研究グループからなり、約60名の学部スタッフと150名の大学院生を擁している。Robotics LaboratoryはComputer science and electrical, computer, and systems engneeringグループに属し、Petros Maragos助教授のもとで信号処理技術を中心とした音声研究がおこなわれている。音声処理を専門とする研究者は数人と少ないが、距離的に近いMITの音声研究グループ（K.N. Stevens 教授）と交流がある。

P. Maragos, A. Potamianos等の手法は、エネルギーオペレータと呼ぶ物理的振動系から導いた概念にもとづいたオペレータを使って音声信号処理をおこなうもので、数年前から始められた独創性の高いものである。この手法を基にして音声のホルマント抽出や音声合成をおこなっている。従来の分析手法から抽出される変量との見かけ上の違いは、きわめて短区間の信号を基に各サンプル点ごとに変量が算出されること、信号の復元性に優れることである。音声再合成のデモでは、代表的な音声分析合成である線形予測法より音質が良いことが直感的に感じられた。ただし、従来の音響音声学的知識と一致しない点もいくつかあり、この辺りの理論的、実験的根拠については今後よく検討していく必要があると考えられる。

音声研究としては、DARPAに参加している機関のものと比べると、きわめて基礎的な範囲にとどまっているといえるが、今日的な状況においてはむしろ貴重である。同所では、他に、関連テーマとしてイメージ処理技術、遠隔操作マニュピレータなどの研究を見学し、意見交換をおこなった。

（3）スタンフォードリサーチ研究所（SRI International）

SRIは、約2000名の研究員を擁し、研究、調査、コンサルティングをおこなっている非営利法人である。その活動は、産業・技術・経営、国際政治・経済の研究、調査、コンサルティング、科学、工学分野の研究、技術開発などで、多くは政府などからの依託業務である。今回はSRIのSpeech Research and Technology Programの部門を訪問した。SRIのこの部門もCMUと同様に、DARPAの音声言語（SLS）プロジェクトの有力メンバであり、高い評価を受けている。ここでは、CMUと同じくフライト情報サービスをタスクとするATISのデモを見学した。また、独自に開発した英語発音訓練システムのデモを見学し、音声認識・理解技術、音声データベースなどの研究動向、問題点について意見交換をした。

Speech Research and Technologyのグループは J. Bernsteinがヘッドで、十数名のスタッフ

がいる。このうち、約3分の1が音声学を専門とする点が特徴である。ATISのデモは、基本的にはCMUと同程度の性能でその差は少ないと思われるが、デモとしての見かけ上はCMUより良くできていた。（昨年度のDARPAワークショップにおけるSRIの報告では、トータルシステムとしてはSRIのシステムの性能がトップであったとなっている）。SRIのシステムに取り入れられている多くの具体的な手法は、CMUなどと比べるとアルゴリズム的な違いはあるが、機能的には大差のないものと考えられる。論文から見る限り、未登録語（辞書にあらかじめ載っていない語）のリジェクションに関するモデルがCMUのシステムにはない機能である。SRIでは、ATISシステムは、話者の協力度、語彙数などでいくつかのレベルに分けられていて、語彙数を1400語まで増やした（通常は1000語）デモシステムもある。

英語発音訓練システムは、学習者が模範となる発音（文）を真似ることによって、その習熟度の向上をはかる方式で、試作段階のものである。学習者の発声に対して、いくつかの項目（たとえば母音、子音、イントネーションなど）に分けて、評価点数が付けらて学習者に知らされる。開発者のGou Kawai氏によれば、ポイントは、どのような分析にもとづいて評価点数をつけるかである。文のセットは、1セット10文で、数種類以上用意してあり、各セットごとにストーリ性があるように作られている。模範発声話者は（今回のテストでは）男性であったが、評価点数をつける場合には、学習者が男／女／子供（年齢によって細分化）であるかを自動的に判断して、評価の規準話者を変えるということである。学習者の発音の評価に力をいれていたので、学習者にどのような提示をするのが適当であるかなどについては、まだ検討の余地を残しているという話であった。

SRIの今後の方向については、少なくともDARPAのプロジェクトが続く限り、SRIの性格からこれに積極的に参加していくと思われる。

第2章　欧　州

第2章　欧　州

2.1　概　要

平成4年11月28日から12月11日にかけて、欧州における音声認識、音声理解および関連技術に関する動向調査を実施した。イギリスでは、ロジカケンブリッジ（Logica Cambridge）、ブリストル大学（University of Bristol）、RSRE（Royal Signals and Rader Establishment）の3機関、ドイツでは、GMD/IPSI（Gesellshaft fur Mathematic und Datenverarbeitung / Integrated Publication and Information Systems Institute）、フィリップスアーヘン研究所の2機関、フランスでは、INRIA（Institute National de Recherche en Informatique et en Automatique）、LIMSI（Laboratoire d'Informatiquepour la Mechanique et les Sciences de l'Ingenieur）の2機関を訪問し、研究討論および情報交換をおこなった。これらの機関は、欧州における人工知能研究の代表的研究機関であり、特に音声認識、音声理解に関する研究分野では多くの先端的な研究がなされている。

研究分野の動向として、音声認識システムの性能が着実に向上していることが感じられた。各研究機関の音声認識システムの認識性能を比較するためのベンチマークテストが、米国DARPAの援助により開催されている音声と自然言語のワークショップでおこなわれてきたが、今回はフィリップスアーヘン研究所とLIMSIの2機関もこれに参加して高い認識性能を達成したとのことであった。

また音声対話を工学的に実現するための対話処理についての研究とともに、言語学的、認知心理学的な側面も含めた、音声対話についての幅広い研究が見られるようになってきている。

音声認識、音声理解技術の応用分野としては、ホームバンキング、列車旅行案内、自動車用ルート案内、オフィスレターの口述等のさまざまな分野で、音声認識、音声理解技術の適用が試みられていた。

欧州での音声認識、音声理解の研究は ESPRIT 計画のもとでもおこなわれてきたが、最近は予算的には伸び悩んでおり、全体的には研究費の確保がなかなか厳しい状況にあるようであった。

2.2　訪問研究機関

（1）ロジカケンブリッジ

ロジカケンブリッジ（Logica Cambridge）では、主にSUNDIAL Project について説明を受けた。

SUNDIAL Projectは 1988年から1993年 8月まで（5年間）ESPRIT 計画のもとでおこなわれている。5カ国からつぎの12機関が参加している。

- UK	Logica (Project Leader)
	University of Surrey
- France	CNET
	CAP SESA Innovation
	IRISA - University of Rennes
- Italy	CSELT
	Saritel
	Politechnico di Torino (associate partner)
- Germany	Daimler Benz
	Siemens
	University of Erlangen
- Sweden	Infovo ((subcontractor to Logica)

プロジェクトの目標は、飛行機や列車での旅行についての情報検索を電話回線を通して音声で対話的におこなうためのプロトタイプの作成である。当初は、語彙2,000語のシステムを目指していたが、現在では最終的には1,000語程度となるものと考えられている。列車情報（独／伊）、フライト情報（英／仏）についてのシステムを現在作成中である。

音声認識、音声合成、対話管理、応答生成などのモジュールは独立しており、全体のアーキテクチャは統一されているが、各モジュールはそれぞれの機関で異なるものを使用している。

Logicaはこのプロジェクトのリーダーであり、英国からはLogicaとサリー大学が参加している。このグループはプロジェクトの初期に、その領域での Wizard-of-OZシミュレーションをおこなって、語彙や文型などの調査をおこなっている。人間－人間の対話と模擬された人間－機械の対話を比較してかなりの差が観察されている。

対話処理の部分は、言語処理、信念の管理、タスク処理、対話管理、応答生成の５つのモジュールで構成されている。それぞれのモジュールは独立して動作し、メッセージ通信によって情報をやり取りする。パーザはユニフィケーションベースで、辞書と知識ベースをDATRと呼ばれる階層的表現で記述している。

他にLogicaでは、call serverと呼ばれる音声認識／音声合成インタフェースを開発している。不特定話者単語音声認識をおこなうもので、ホームバンキングなどを応用として想定している。

（２）ブリストル大学

ブリストル大学のWrightのグループは、統計的なLR構文解析手法やあいまいさの残る仮説

をまとめて保持することによる高速なN-best探索手法などのさまざまな研究をおこなってきた。音声認識については他の機関で開発されたものを利用しており、言語処理に研究の重点を置いている。最近の仕事の1つとして、robustな処理を目指して、単語bigramを用いたプロセスと文脈自由文法を用いたプロセスを並列に動作させ、それぞれから得られたN-best仮説を統合する手法についての研究がある。

現在は、マルコフモデルに見られる逐次的・データ駆動的な手法と、文脈自由文法を用いた構文解析に見られる階層的・ルールベースの手法の統合を目指しており、部分文パーザと非終端記号のbigramの組み合わせによるrobustな構文解析手法や、テキストの文脈に依存した確率な言語モデルについての研究をおこなっている。

（3）RSRE

RSRE（Royal Signals and Radar Establishment）は、2年程前にその名前をDRA（Defense Research Agency）と変えた。信号処理分野についての長い歴史をもち、多くの有名人を出している。また Dr. J. Bridle は最近、Dragon UK に移った。

音声研究ユニットは、パターン認識、音声パターン処理、技術評価、応用のグループに分かれていて、15人くらいの規模である。2年程前にARMADAという音声認識システムを開発している。語彙500語で、言語的制約なしの不特定話者連続音声認識をおこなうものである。トランスピュータを10個用いて実時間処理を実現している。有効性が知られているいくつかの技術を集めて高性能を達成しているという印象を受けた。

これに関連してAurixと呼ばれる音声認識システム開発キットを用意している。これはさまざまな応用システム開発者が音声認識システムについて知ることなしに、語彙や文法などのシステムの仕様をPC上のツールで決定してメモリカードに書き込み、それを認識システムが用いるというものである。

またWizard-of-OZシミュレーションによって、人間－機械対話について調べている。応答はシステムでなく、あらかじめ決められたプロトコルにしたがって人間が話し、その声を人工的に変えて合成音らしくしたものをユーザに聞かせている。人間が応答することによって、制約の少ない、より自由な対話を対象としようとしている。電話を通して自動車用のルート案内をするもので、最短ルートなどはPC上で動作する既存のパッケージを利用している。

その他、N-best探索、決定木クラスタリングなどについて最近発表している。

（4）GMD/IPSI

GMD（Gesellschaft fur Mathematik und Datenverarbeitung mbH）「ドイツ国立情報処理研究所」は、情報科学技術分野における基礎研究を含めた幅広い研究活動を促進するために、

1968年に設立された研究機関である。財政的には、連邦政府が90%を、所在地の州が10%を拠出している。主務官庁は連邦研究技術省（BMFT）である。スタッフ数は約1,400人、内約900人が研究に従事している。年間予算は約120億円である。

IPSI（Integrated Publication and Information Systems Institute）「統合出版・情報システム研究所」はGMDの８つの研究所の１つである。スタッフ数は110人で2/3が研究者である。印刷物および電子媒体（ビデオ、アニメーション、グラフィクス、音響、音声）を含めたマルチメディア出版システムが重要な研究対象である。

訪問したDr. Batemanらのグループはマルチメディア知識処理のグループで、主に自然言語処理の解析と生成について研究している。スタッフは解析が６名、生成が４名とのことであった。

特定の応用に依存しない一般的な自然言語処理を実現することと、文脈情報の利用が主な目標である。文の構造、ローカルなテーマ、照応構造、テキストの概念構造、背景知識を５層の階層構造で表現し、モジュール化を図っている。生成は、ドイツ語、英語、オランダ語を対象としている。辞書の語義文からの概念構造の獲得といったことも研究の対象としている。

（５）Philips Aachen

フィリップスアーヘン研究所は、物理学、化学、電子工学分野の研究をおこなう研究所で、470人程度の規模である。ここにはDr. NeyをHeadとして15人程度の規模の音声認識のグループがある。小語彙、対話システム、ディクテーションを対象としているが、現在はディクテーションに重点を置いているとのことであった。

小語彙のシステムではノイズ除去、音素文脈依存モデルの状態を単位とするLDA（Linear Discriminant Analysis）によるrobustな音声認識について研究している。対話システムは、電話回線を通して列車の時刻表についての案内をおこなうものである。デモシステムでは、語彙数2,500語の不特定話者の連続音声を４台のWSで分散処理して、実時間で認識している。

ディクテーションシステムは、オフィスレターを想定した12,000語の語彙の特定話者連続音声認識システムである。PC+DSPboardで構成されたデモシステムでは、実時間の２倍程度の処理時間であった。認識システムは、LDAをフロントエンドとした連続分布HMMで、Viterbi Training、音素文脈依存モデル（単語内のみ）の使用といった特徴がある。

統計的言語モデルについては、non-linear smoothing、キャッシュ、単語に付与するカテゴリーの自動学習といったテーマについて研究している。

このグループは、1992年９月にDAPRAのResource Management Benchmark 評価に参加し、特定話者、不特定話者で高い認識性能を示して、その実力の程を示した。

BMFTのプロジェクトとしては、ASL（Architecture for Speech Language）が２年前から

スタートしている。また現在 VerbMobilプロジェクトが提案中で、ドイツ語、英語、日本語のポータブルな音声翻訳システムを目指す8年間のプロジェクトであるとのことであった。

（6）CRIN/INRIA

CRIN（Centre de Recherche en Informatique de Nancy）は、CNRS（Centre National de la Recherche Scientifique）傘下の研究所であると同時にINRIA（Institute National de Recherche en Informatique et enAutomatique）の1研究所でもあり、また大学院の学生の指導もするという複合的な組織である。現在、約250名の規模である。

音響・音声学的な知識をもとにコンテキストごとに異なる音韻の特徴をエキスパートシステムによって記述しようとするルールベースの音韻認識、高次からのフィードバックを利用したノイズ環境に強い聴覚モデル、主成分分析を組み合わせた topological map による音韻認識と単語認識などのいくつかのテーマについて研究している。

また語彙数37,000語のディクテーションシステムについて研究している。統計的言語モデルを用いた不特定話者連続音声認識システムである。40万語のテキストから単語カテゴリーのtri-gramとカテゴリーの定義を自動的に学習しようとしている。

（7）LIMSI

LIMSI（Laboratoire d'Informatique pour la Mechanique et les Sciencesde l'Ingenieur）は、CNRS傘下の研究所である。Human-Machine Communication Departmentは約110人の規模であり、Speech Communication Group、Language& Cognition Group、Non-verbal Communication Group、Human Cognition teamの4つのグループから構成されている。Human Cognition teamは今年からスタートしたグループである。

音声関係では、音声分析、音声合成、評価、話者の多様性、音声認識、対話構造、音声言語モデル、コネクショニストシステム、マルチモーダルインタフェースなどのテーマについて、認知科学的な研究から、開発したVLSIを用いた並列処理システムまで、幅広くカバーしている。

Speech Communication Group では、音声タイプライタと対話システムをゴールとして、幅広い研究をおこなっている。このグループは、波形レベルの詳細な分析、話者や話し方のスタイルの多様性についての対話システムを用いた研究、大規模な音声データベースを用いた音声認識システムの評価などが行われている。

LIMSI もフィリップスと同じく、1992年 9月にDAPRAのResource Management Benchmark 評価に参加し、とくに不特定話者で高い認識性能を示している。この音声認識システムは、連続分布HMMをベースとし、流暢な音声に見られる変形をルール化して学習時と認識時

に使用している点に特徴がある。

対話構造についての研究では、早くから (1984年～) Wizard of OZ シミュレーションをおこなって、言語学的な側面について調べている。航空管制システムは、実際に訓練をおこなう教育機関で使用される予定とのことであった。

言語モデルの研究は、統計的手法を中心としておこなわれており、文法的なカテゴリーの自動学習などについて最近は研究がおこなわれている。コネクショニストシステムの研究もおこなっており、音声サンプルのみから単語カテゴリー集合を学習しようとする研究などがある。

この他に、Non-verbal Communication Group ではビジョン、音声などを含むさまざまなマルチモーダル対話システムについて研究している。

全体に、先端的な研究分野をよくカバーした非常に活発な研究活動をおこなっているという印象であった。

第 3 章　ICSLP 92

第３章　ICSLP 92（音声言語処理に関する国際会議 - 92）

３．１　会議の概要

この国際会議は、８学会（米国音響学会、日本音響学会、カナダ音響学会、カナダ言語学会、ヨーロッパ音声通信学会、IEEE信号処理学会、国際音声学会、米国言語学会）が提携して、カナダ・アルバータ大学の主宰で、アルバータ州バンフのBanff Springs Hotelにおいて、1992年10月12日～16日の期間開催された。この分野の国際会議としては世界最大規模である。今回が第２回目であり、第１回は1990年に神戸で開催され、それ以後、隔年ごとに開催されることで合意された。次回は1994年横浜で開催されることが決っている。

今回の会議では、合計442件の研究発表があり、このうち267件が口頭発表、残りがポスターでの発表であった。また、４件のplenary lectureがあった。全体のプログラムは表３．１-１に示す通りで、口頭発表は４つのセッションが並行しておこなわれ、さらにポスターセッションが１つ設けられているという形式であった。オープニングセレモニーでは、UCLA（カリフォルニア大学ロサンジェルス校）の著名な音声学者 Peter Ladefogedが"Knowing enough to analyze spoken languages"と題して講演した。

表３．１-１　ICSLP 92　PROGRAM SCHEDULE

				1	2	3	4	Posters
TUESDAY	Morning	first	0830-1010	OPENING & PLENARY				
		second	1030-1210	Word Spotting in ASR	Speech Coding	Speech Production: Coarticulation	Auditory Models	
	Afternoon	first	1400-1540	Recognition of Telephone Speech	Text-to-Speech Synthesis I	Voice Source Characteristics	Speech Perception: Higher Order Processes I	Speech Production, Perception and Analysis
		second	1600-1740	Speaker-Independent Word Recognition	Text-to-Speech Synthesis II	Human Factors	Speech Perception: Higher Order Processes II	
WEDNESDAY	Morning	first	0830-1010	Continuous Speech Recognition I	Natural Language Processing and Speech Understanding I	Language Learning and Acquisition I	Speech Perception: Higher Order Processes III	Synthesis
		second	1030-1210	Continuous Speech Recognition II	Natural Language Processing and Speech Understanding II	Language Learning and Acquisition II	Prosody: the Phrase and Beyond I	
	Afternoon	first	1400-1540	Speaker Adaptation	Natural Language Processing and Speech Understanding III	Pronunciation Training	Prosody: the Phrase and Beyond II	
		second	1600-1740	PLENARY				
THURSDAY	Morning	first	0830-1010	Self Organizing Systems in ASR I	Synthesis I	Speech Perception: Phonetic Processes I	Speech Enhancement	Linguistic Phonetics
		second	1030-1210	Self Organizing Systems in ASR II	Synthesis II	Speech Perception: Phonetic Processes II	Speaker Recognition	
	Afternoon	first	1400-1540	Hidden Markov Models	Dialogue	Prosody and Phonology I	Speaking Styles I	Perception and Production
		second	1600-1740	ASR in Noise	Dialogues and Applications	Prosody and Phonology II	Speaking Styles II	
FRIDAY	Morning	first	0830-1010	Sampling of Speech Production I	Labelling of Speech I	Speech Production: Models & Theories I	Miscellaneous	Speech Recognition
		second	1030-1210	Sampling of Speech Production II	Labelling of Speech II	Speech Production: Models & Theories I	Speech Data Bases	
	Afternoon	first	1400-1540	Stochastic ASR	Analysis of Disfluencies in Speech I	Linguistic Phonetics: Reduction	Hearing/Speech Impaired I	Analysis/Systems
		second	1600-1740	Unit-Based ASR	Analysis of Disfluencies in Speech II	Language and Dialect Characterization	Hearing/Speech Impaired II	

会議参加者の人数はおおむね600人程度と見積られ、全体を通じて盛況であった。研究発表件数を国別に見ると表3.1-2のようになり、合計で27か国からの発表があったことになる。したがって、これらの国を中心に第一線の研究者が参加したものと見込まれる。参加者の中には、音声研究の分野で1950年代より活躍しているGunner FantやK. N. Stevensなど著名な研究者多数が含まれ、彼等自身が研究発表をおこなっていた。

表3.1-2　国別発表件数

米国	115	日本	89	カナダ	37	フランス	30
英国	36	ｵﾗﾝﾀﾞ	24	ｽｳｪｰﾃﾞﾝ	16	ドイツ	17
イタリー	14	中国	14	スペイン	10	ﾃﾞﾝﾏｰｸ	9
ﾌｨﾝﾗﾝﾄﾞ	6	台湾	4	ベルギー	4	韓国	4
インド	2	ﾉﾙｳｪｰ	2	その他9ｶ国	1		

分野ごとの研究発表の件数をまとめるとおよそ表3.1-3に示すようになる。この表からも分かるように、音声認識、理解、符合化、合成などのいわゆる工学的分野にとどまらず、音声知覚、生成、言語音声学などの分野の発表も数多くおこなわれた。これらの中から今回の調査目的に合わせて選んだ分野ごとの概要を次節に述べる。

表3.1-3　分野別の発表件数（概数）

（口頭発表）	
1）自動認識、理解（単語、連続音声、話者適応等）	75
2）符合化、合成	45
3）データベース、ラベリング、方言	15
4）発話解析（Speaking Style, Disfluencies） 対話解析	45
5）話者認識	5
6）韻律、音韻論、音声学	25
7）言語習得、獲得	15
8）音声知覚、生成	50
（ポスターセッション）	
1）分析、認識	58
2）合成	28
3）言語音声学	36
4）知覚、生成	61

3.2 分野毎の研究発表概要

（1）音声認識関連

音韻認識に関連した手法としては、ニューラルネットワーク、コホネン（Kohonen）の自己組織化写像、学習ベクトル量子化などとHMMを結び付けたものが多く見られた。

単語認識や連続音声認識に関連した研究発表では、従来のシステムに比べ、一段と大語彙化が目立った。たとえば、INRS-Telecommunications（Canada）では、語彙数6万語の連続音声システムの構築のための高速検索アルゴリズムや、音韻HMMを区切りの無い連続音声から訓練（学習）する方法などを開発している。ほとんどの認識システムは、音韻HMMをベースとしたシステムでHMMの訓練方法や辞書びきのアルゴリズムなどについての発表が多い。

電話回線を利用した認識システムの発表が5件ほどあったが、多くは大語彙の単語認識をベースとする方式である。この中で、M. Lennig (Bell Northern Research) 等の発表では、株式市況に関する問い合わせをタスクとするシステムを会場で実演して見せた。話者不特定のシステムで、語彙は会社名など1,561単語であるが、音韻単位のモデルをベースにしているので語彙セットの変更が容易である。

話者適応化手法に関する発表は、事実上、日本からのみであり、ベクトル量子化コードの写像にもとづいた方式が一段落したという状況がみられる。

（2）音声理解、対話、自然言語処理関連

ここで発表された研究の多くは、それぞれの国で知られる大規模プロジェクトの下でおこなわれている。すなわち、米国ではDARPAの音声言語システム（SLS）プロジェクトのタスクであるATIS（フライト情報案内）、ECではESPRIT（Ⅱ）におけるSUNDIAL、SUNSTARなどのプロジェクト、日本ではATRのプロジェクトにおける研究が発表された。

手法的には、確率言語モデルによる処理を目指したものが3件、文脈自由文法にもとづくものが3件、弛緩法を導入した方式が1件、また、タスクに依存したドメイン限定テンプレートとシンタクティックテンプレートの統合を扱った方式があった。

対話システムの使い勝手の評価を扱った発表として、いわゆるWizard of OZ（音声認識システムを人間が代行する方式。ただし、利用者にはその事実が見えないようにしておく）を用いて認識精度と利用者の評点の関係を調べたものがあり、認識精度の高いことが評点の良さに大きく寄与していると述べている。

（3）音声データベース、ラベリング

音声データベースに関しては、電話系を通した大語彙データベースおよび多言語データコーパス（R. Cole, et al, Oregon G. Institute）、経済誌（Wall Street Journal）の連続音声デー

タベース（D.B. Paul, et. al, MIT等、これはDARPAのSLSプロジェクトの一環で、大語彙連続音声のディクテーションを指向したもの）についての発表があった。

音声データのラベリングの問題に関連して、従来のような音声的音素的単位へのラベリングに加えて、韻律に関するラベリング方式の提案があった（K. Silverman, et. al, NYNEX）。これは、米国語の韻律に関するラベリングの標準化を目指したもので注目される。

（4）発話様式、会話音声、自由発話などの解析

これらに関連した研究発表が増加したことが今回の会議の特徴の1つであったといえる。まず、読み上げ音声と自由発話音声の比較、この中には、音声学的比較、ピッチの比較、単語の使用についての比較などがあった。また自由発話音声の中でもカジュアル（causual）とケアフル（careful）の比較などの解析結果が発表された（E. Blauw, et. al; M. Eskenazi, et. al など5件）。

情緒表現の音響的特性の検討、およびその音声合成への適用などついての発表があり、必ずしも明解な知見とはいかないが、今後に残されたテーマが扱われるようになってきたといえる。Hesitation（躊躇）の分析や笑い声の音響的分析、自由発話におけるいわゆるfalse startの分析などがDisfluencies in Speechのセッションとして8件ほどあった。

関連した話題として、電話番号案内におけるオペレータと顧客の対話の解析、マン-マシン対話（旅行案内）におけるエラーメッセージの効果に関する解析などの発表があった。

（5）音声合成、その他

音声合成に関連した研究発表は、テキスト-音声変換、音韻合成、また、プロソディーや音源、音声生成などの一部にもあって、総数としてはかなり多かった。ここでは、（EC統合に関連してか）、多言語の音声合成とその評価の発表があったことだけ記しておく。

その他、音声知覚や聴覚モデルのセッション、言語獲得のセッションなどにも興味深い研究発表があったが、今回の調査対象と若干離れるので、ここでは省略する。

3.3 音声データベースと評価の国際協調に関するワークショップ

標記のテーマに関するワークショップがICSLP 92の会議に先立ち、10月12日に開催された。この会議は、第1回目がICSLP 90に伴って神戸で開催され、その後昨年、EUROSPEECHのとき第2回目が開催された。この時、音声合成評価WG、音声認識WG、データベースWGが設けられ、全体の世話人として、フランスLIMSIのJoseph J. Marianiがあたることになった。今回が第3回目で、Cocosda Meetingという名称が付けられている。

今回の参加国は十数カ国で、参加者数はおよそ100名弱であった。会議では、まず米国、日

本、EC（英、仏、独、伊、蘭など）、韓国などの音声データベースや音声関連プロジェクトについての紹介があり、次に各ワーキンググループの活動状況報告などがあった。実質的な活動は、これからという状況である。

今回の会議で、決まったこととしては、数字音声の5000人分のデータを電話系を通して各国で収集し、相互に交換するというものである。この提案には12カ国（米、日、英、仏、独、伊、蘭等）が賛同した。来年、開催されるEUROSPEECH（於：ベルリン）の際にその進捗状況を報告するということになっている。また、電子メールのグループリストを作成し、情報交換を密にすることも確認された。

資　料　編

A.　「研究用連続音声データベース」利用ツール

第1章　概　要

第1章　概　要

「研究用連続音声データベース」は、(社)日本音響学会の「連続音声データベース調査委員会」での検討結果にもとづいて、実際に国内の大学、国立研究所、企業等の１７機関が協力して音声データを分担収録したものである。これらのデータはCD-ROM ７巻に収納され、現在、(財)日本情報処理開発協会から音声研究者等を対象に広く配布されている。

CD-ROMのデータフォーマットはISO9660の規格にしたがっているが、これらのCD-ROMを入手して、ワークステーションやパソコンで音声データを読み出して実際に使用するためには、インタフェースソフトウェアを作成する必要がある。

また、初めての人が音声の研究を始めようとする場合には、CD-ROM内の音声データの表示やラベリング、分析等のソフトウェアも作らなければならない。

そこで、本委員会では、「研究用連続音声データベース（CD-ROM ７巻）」を、入手してすぐに使用できるようなツールと、音声研究を始めようとする人が必ず作らなければならない研究ツールについて検討をおこない、これらのツールを「研究用連続音声データベース」利用ツールとして作成した。以下にその仕様、利用手順等について説明する。

なお、この利用ツール内のＨＭＭ関係プログラムについては豊橋技術科学大学の中川聖一教授の研究室から提供されたプログラムをベースにして改定作成した。

また、本利用ツールは、(財)日本情報処理開発協会より提供（CD-ROM版）を予定している。

第２章　研究用連続音声データベース

第２章　研究用連続音声データベース

２．１　はじめに

研究用連続音声データベース全７巻のCD-ROM（Vol.1～Vol.7）は次の３つに分類される。

(1) Vol.1～Vol.3 にはATR 音素バランス文（503文）を64名（男30名、女34名）の話者が発声した約9,600文の音声波形とテキストが収録されている。

(2) Vol.4～Vol.6 には道案内、観光案内、問い合わせ等各種案内をタスクとした模擬対話文（1,027文）を36名（男女各18名）の話者が発声した12,474文の音声波形と発声テキストが収録されている。さらにVol.6 には読み上げの連続音声、模擬対話の書き起こしテキスト、報告書が含まれている。

(3) Vol.7 には道案内、観光案内、問い合わせ等各種案内をタスクとした模擬対話（40対話）を各々２名の話者が発声した音声波形と発声テキストが収録されている。

音声は、16kHz、16ビット（一部１ビット）でディジタル化されている。本データは(社)日本音響学会の連続音声データベース調査委員会（1990年７月設置）で検討し、17機関が協力して分担収録したものである。しがって、録音やAD変換（低域炉波器特性(6kHzまで平坦、8kHzで -40dBが基準）、量子化ビット数）特性は必ずしも統一されてはいない。

２．２　音声データベースの内容

２．２．１　ATR音素バランス　503文（Vol.1～Vol.3）

2音素連鎖（CV(120種),VC(227種),VV(55種),計402種）および一部の3音素連鎖（CVC(69種、ただしCは無声破裂音、無声摩擦音）、CVC（18種、ただしCは鼻子音）、VCV（136種、ただしCは半母音),計223種）について、これらの音素環境が同じ割合で含まれるようにするため、これらの出現を独立と見なして情報エントロピーを定義した。文の母集団としては、新聞、雑誌、小説、手紙、教科書等から無作為に10,196文を抽出し、これらを上述の基準にしたがって最適化手順により503文を選んだ。これは50文を１セットとして音素バランスがとれるよう配置されている。

ATR音素バランス503文には10のサブセットがある。A,B,C,...,I は各50文からなり、Jは53文からなる。

Aセットはすべての話者が発声し、他のセットは、原則として男女各６名 計12名が発声している。

音声データはすべて検聴し、そのチェックリストもファイルに格納されている。また、チェックの結果、再録音をした場合、修正前のデータ（言い淀み等を含む）も、補助データとして、

別ディレクトリに格納されている。

2.2.2 各種案内模擬対話 (Vol.4～Vol.6)

(1) 音声データ

道案内(5種)、観光案内(7種)、音楽会案内(2種)、パスポート取得問い合わせ、スキーツアー案内の計16対話からなる。これは、タスクを決めて2人の話者が模擬対話をおこない、その音声を文字に書き起こして間投詞や言い間違い等を除いた文に直し、それを改めて一人の話者が読み上げたものである。

道案内と観光案内の一部(金沢案内)と音楽会案内については、助詞の脱落・省略や文の倒置をそのままにしたものと、その脱落・省略を補い、倒置を修正して丁寧な表現に修正したものの両方が収録されている。

模擬対話文の話者36名中8名(男女各4名)は、ATR音素バランス503文(Vol.1～Vol.3)の話者と同一である。それ以外の話者は重複していない。全体は16のセット(k, l, m,..., z)に分かれており、各セットは30～141文からなる。各セットは10～18名の話者が発声し、各話者は5～9セットを発声している。

音声データはすべて検聴し、そのチェックリストもファイルに格納されている。またチェックの結果、言い淀みや言い誤りが見つかった場合は再録音をしている。

(2) 連続音声データベース関連情報(報告書)

Vol.6には、CD-ROM Vol.1～Vol.6の連続音声データベースとテキストに関連した資料が含まれている。連続音声データベースに関するサーベイ、模擬対話と書き起こしテキストに関する諸問題(タスクの定義、表記法、書き起こしテキストから読み上げテキストを作成する際の基準など)、連続音声データベースの仕様などが記述されている。

2.2.3 各種案内模擬対話 (Vol.7)

道案内、観光案内、音楽会案内、パスポート取得問い合わせ、スキーツアー案内等の40対話からなる。これは、タスクを決めて2人の話者が模擬対話をおこなったもので、テキストはその音声を文字に書き起こしたものである。

2.3 CD-ROMフォーマット

CD-ROMは、ISO9660フォーマット、音声データはIEEE形式(2バイト/1サンプル、上位バイト,下位バイトの順)で格納されている。波形ファイルにはヘッダーはなく、音声データのみが格納されている。

2.4 ディレクトリ構造

（1）ATR音素バランス連続音声データベース(Vol.1～Vol.3)

①CD-ROM Vol.1 のディレクトリ構造

```
vol1-+-doc-+-readme. (this file)
   ¦   ¦
   ¦   +-kaisetsu.jis (general instruction in Japanese, JIS-code)
   ¦   ¦
   ¦   +-kaisetsu.sjs (general instruction in Japanese, Shift-JIS)
   ¦   ¦
   ¦   +-instruct.    (general instruction in English)
   ¦   ¦
   ¦   +-text-+-romaji-+-readme. (instruction of transcription)
   ¦   ¦     ¦       ¦
   ¦   ¦     ¦       +-a.rom (ro-maji transcription of "a" set)
   ¦   ¦     ¦       :
   ¦   ¦     ¦       +-j.rom (ro-maji transcription of "j" set)
   ¦   ¦     ¦
   ¦   ¦     +-kanji--+-a.kan (kanji transcription of "a" set)
   ¦   ¦     ¦       :
   ¦   ¦     ¦       +-j.kan (kanji transcription of "j" set)
   ¦   ¦     ¦
   ¦   ¦     +-kana---+-a.hir (kana transcription of "a" set)
   ¦   ¦             :
   ¦   ¦             +-j.hir (kana transcription of "j" set)
   ¦   ¦
   ¦   +-speakers-+-readme. (information about speakers)
   ¦              ¦
   ¦              +-checklst. (checklist of speech wave by listening)
   ¦
   +-dat-+-tsu0001-+-a-+-a01.ad (speech wave, a01 of speaker tsu0001)
         ¦         ¦ :
         ¦         ¦ +-a50.ad (speech wave, a50 of speaker tsu0001)
         ¦         ¦
```

```
¦        +-b-+-b01.ad (speech wave, b01 of speaker tsu0001)
¦        ¦  :
¦        ¦  +-b50.ad (speech wave, b50 of speaker tsu0001)
¦        ¦
¦        +-c-+-c01.ad (speech wave, c01 of speaker tsu0001)
¦           :
¦           +-c50.ad (speech wave, c50 of speaker tsu0001)
:
:
:
```

②CD-ROM Vol.2 のディレクトリ構造

vol2-+-doc-※（Vol.1 の doc 以下と同じ）

```
¦
+-dat-+-etl0001-+-a-+-a01.ad (speech wave, a01 of speaker etl0001)
¦        ¦  :
¦        ¦  +-a50.ad (speech wave, a50 of speaker etl0001)
¦        ¦
¦        +-c-+-c01.ad (speech wave, c01 of speaker etl0001)
¦        ¦  :
¦        ¦  +-c50.ad (speech wave, c50 of speaker etl0001)
¦        ¦
¦        +-d-+-d01.ad (speech wave, d01 of speaker etl0001)
¦           :
¦           +-d50.ad (speech wave, d50 of speaker etl0001)
:
:
:
```

③CD-ROM Vol.3 のディレクトリ構造

vol3-+-doc-※（Vol.1 の doc 以下と同じ）

```
¦
+-dat-+-fuj0001-+-a-+-a01.ad (speech wave, a01 of speaker fuj0001)
```

```
¦       ¦  :
¦       ¦  +-a50.ad (speech wave, a50 of speaker fuj0001)
¦       ¦
¦       +-b-+-b01.ad (speech wave, b01 of speaker fuj0001)
¦       ¦  :
¦       ¦  +-b50.ad (speech wave, b50 of speaker fuj0001)
¦       ¦
¦       +-c-+-c01.ad (speech wave, c01 of speaker fuj0001)
¦          :
¦          +-c50.ad (speech wave, c50 of speaker fuj0001)
:
:
:
```

（2）各種案内模擬対話文音声データベース（Vol.4～Vol.6）

①CD-ROM Vol.4 のディレクトリ構造

```
vol4-+--doc--+-readme. (this file)
   ¦     ¦
   ¦     +-kaisetsu.jis (general instruction in Japanese, JIS-code)
   ¦     ¦
   ¦     +-kaisetsu.sjs (general instruction in Japanese, Shift-JIS)
   ¦     ¦
   ¦     +-instruct.    (general instruction in English)
   ¦     ¦
   ¦     +-text-+-kana---+-k.hir (kana transcription of "k" sentences)
   ¦     ¦    ¦      :
   ¦     ¦    ¦
   ¦     ¦    +-kanji--+-k.kan (kanji transcription of "k" sentences)
   ¦     ¦    ¦      :
   ¦     ¦    ¦
   ¦     ¦    +-romaji-+-readme. (instruction of transcription)
   ¦     ¦           +-k.rom (ro-maji transcription of "k" sentences)
```

```
¦     ¦            :
¦     ¦
¦     +-speakers-+-readme. (information about speakers)
¦                +-checklst. (checklist of speech wave by listening)
¦
+--dat--+-can0002-+-k-+-k001.ad
¦     :       :  :
¦
+-kana--+-can0002-+-k-+-k001.hir
¦     :       :  :
¦
+-kanji-+-can0002-+-k-+-k001.kan
¦     :       :  :
¦
+-romaji+-can0002-+-k-+-k001.rom
      :       :  :
```

/vol4/dat/can0002/k/k001.ad

is the speech wave file of k001 sentence by speaker can0002

/vol4/romaji/can0002/k/k001.rom

is the ro-maji transcription of k001 sentence by speaker can0002

/vol4/kana/can0002/k/k001.hir

is the kana transcription of k001 sentence by speaker can0002

/vol4/kanji/can0002/k/k001.kan

is the kanji transcription of k001 sentence by speaker can0002

②CD-ROM Vol.5 のディレクトリ構造

vol5-+-doc-※（Vol.4 の doc 以下と同じ）

```
¦
+--dat--+-mac0003-+-o-+-o001.ad
¦     :       :  :
¦
+-kana--+-mac0003-+-o-+-o001.hir
```

```
|      :        :  :
|
+-kanji-+-mac0003-+-o-+-o001.kan
|      :        :  :
|
+-romaji+-mac0003-+-o-+-o001.rom
       :        :  :
```

/vol5/dat/mac0003/o/o001.ad

is the speech wave file of o001 sentence by speaker mac0003

/vol5/romaji/mac0003/o/o001.rom

is the ro-maji transcription of o001 sentence by speaker mac0003

/vol5/kana/mac0003/o/o001.hir

is the kana transcription of o001 sentence by speaker mac0003

/vol5/kanji/mac0003/o/o001.kan

is the kanji transcription of o001 sentence by speaker mac0003

③CD-ROM Vol.6 のディレクトリ構造

vol6-+-doc-※（Vol.4 の doc 以下と同じ）

```
|
+-trnscrpt-+-readme. (information about the trancsribed texts)
|        +-crl-+-crl0001d. (trancsribed texts)
|        :
|
+-report+-mokuji. (table of contents of report)
|     +-honbun-+-ch1. (Introduction)
|     |        +-ch2. (Survey of Continuous Speech Database)
|     |        +-ch3. (Simulated Dialogue Database)
|     |        +-ch4. (Texts for Read Speech)
|     |        +-ch5. (Continuous Speech Data)
|     |        +-ch6. (Future Subjects)
|     +-huroku-+-apena. (Morphological Analysis)
|              +-apenb. (Text Database for Turn Taking Research)
```

```
¦           +-apenc. (Transcription Character Set)
¦           +-apend. (Guide for Collection of Speech Data)
¦           +-apene. (Speakers and Texts)
¦           +-apenf. (Checklist of Speech Data)
¦           +-apeng. (Microphones and Filters for Recording)
¦           +-pbs-+-kana-+-a.hir ("A set" of ATR 503 Phonetically
¦           ¦   ¦    :        Balanced Sentences in kana)
¦           ¦   +-kanji-+-a.kan ("A set" of ATR 503 Phonetically
¦           ¦   ¦    :        Balances Sentences in kanji)
¦           ¦   +-romaji-+-a.rom ("A set" of ATR 503 Phonetically
¦           ¦        :        Balances Sentences in romaji)
¦           +-vgts-+-kana-+-k.hir ("K set" of Various Guide Task
¦           ¦   ¦    :        Sentences in kana)
¦           ¦   +-kanji-+-k.kan ("K set" of Various Guide Task
¦           ¦   ¦    :        Sentences in kanji)
¦           ¦   +-romaji-+-k.rom ("K set" of Various Guide Task
¦           ¦        :        Sentences in romaji)
¦           +-tagged-+-readme. (information about tagged texts)
¦                 +-etl0008r. (tagged texts)
¦                 :
+--dat--+-sha0001-+-o-+-o001.ad
¦     :       :  :
¦
+-kana--+-sha0001-+-o-+-o001.hir
¦     :       :  :
¦
+-kanji-+-sha0001-+-o-+-o001.kan
¦     :       :  :
¦
+-romaji+-sha0001-+-o-+-o001.rom
      :       :  :
```

/vol6/dat/sha0001/o/o001.ad

is the speech wave file of o001 sentence by speaker sha0001

/vol6/romaji/sha0001/o/o001.rom

is the ro-maji transcription of o001 sentence by speaker sha0001

/vol6/kana/sha0001/o/o001.hir

is the kana transcription of o001 sentence by speaker sha0001

/vol6/kanji/sha0001/o/o001.kan

is the kanji transcription of o001 sentence by speaker sha0001

第３章　ハードウェア構成およびツールの機能

第３章　ハードウェア構成およびツールの機能

３．１　ハードウェア構成

本ツールを利用する場合の基本的なハードウェア構成は、以下のとおりである。

(1) 本体: Sun SPARC station シリーズ

(2) ＯＳ: SunOS 4.1.2

(3) ウィンドウシステム: OpenWare/Motif ver2 および日本語OpenWindows 2.0.1

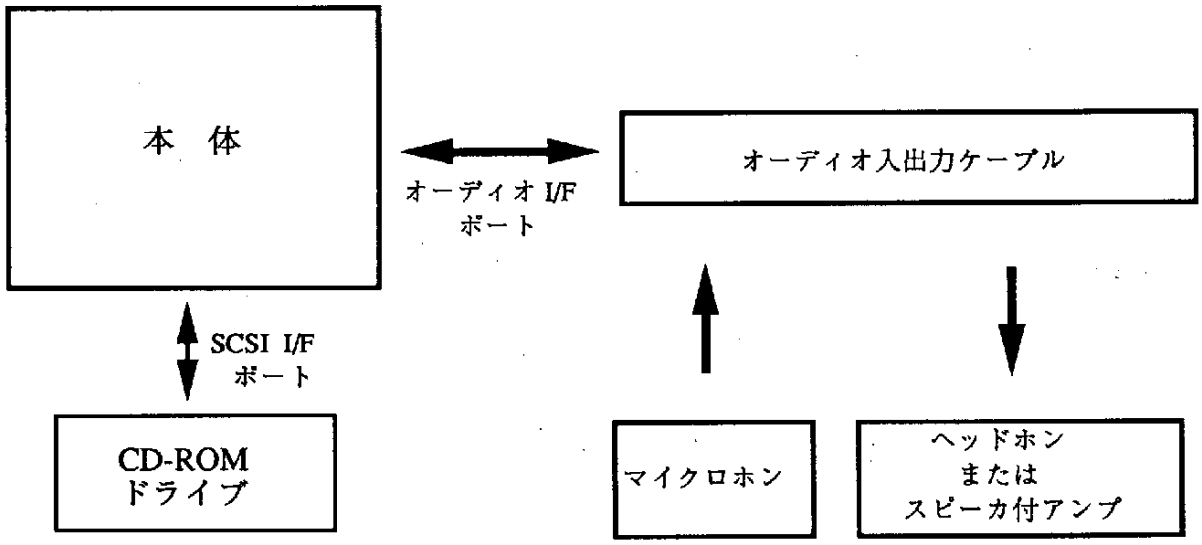


図３．１-１　ハードウェア構成

３．２　ツールの機能

本ツールは、①研究用連続音声データベース（CD-ROM Vol.1～Vol.7）のハンドリング機能、②音声データ研究のための基本的プログラム（関数ライブラリ）、③HMMにもとづく音声認識関係プログラムを提供している。

なお、これらのプログラムのソースは全てSunの標準Ｃ言語で記述している。

3.2.1 音声データベースハンドリング機能

（1）音声の録音・再生機能

① 音声の録音：最大5分（ただしシステム構成により制限される場合がある）

② 音声の再生、繰り返し再生（ループ再生）

③ 再生範囲の設定

④ 録音・再生の一時停止とその解除

⑤ 録音・再生レベルの調整

⑥ 録音・再生時間と波形のモニター

（2）ファイル操作

① サンプリングレートの変換

② 保存範囲の設定

③ 既存ファイルへのアペンド保存

（3）波形編集・表示機能

① 指定領域のコピー、削除、ペースト

② 指定領域の拡大表示

（4）スペクトル分析機能

① 指定領域のスペクトログラム分析と結果の表示

② スペクトログラム結果表示の濃度調整

③ ピークピッキング結果の表示

（5）ラベリング編集機能

① ラベル境界の挿入・削除

② ラベル境界位置の量子化

③ ラベル境界およびラベル記号の異なる層間でのコピー

④ ラベル記号の編集

（6）コマンド実行機能

① 画面のハードコピーなどの UNIX コマンドの実行

3.2.2 関数ライブラリ

音声データの操作、分析のための以下の関数をライブラリとして用意している。

関数名	機　能
set_window()	音声データに指定した時間窓を乗じて短区間を切り出す。時間窓は、矩形窓、ハニング窓、ハミング窓、ブラックマン窓、三角（Bartlett）窓の5種類の中から指定する。
durb()	Durbin法により音声信号から自己相関関数値、線形予測係数、PARCOR係数を求める。
lp2cep()	線形予測係数からLPC cepstrum係数を求める。
cep2mel()	LPC cepstrum係数から mel cepstrum係数を求める。
lp2power()	線形予測係数からパワースペクトルを求める。
fft()	高速フーリエ変換。
power2pk()	パワースペクトルからピークピッキング法によりホルマントを求める。

3.2.3 HMM関係プログラム

HMMにもとづく音声認識プログラムとして、以下のプログラムを用意している。

プログラム名	機　能
TR_CMHMM	学習プログラム
DP_CMHMM	単音節認識プログラム（孤立音声の認識用HMM）
ONDP_CMHMM	O(n)DP法による連続音節の認識
OPDP_CMHMM	オートマトン制御On ePass DP法による連続単語認識

※以上はメルケプストラムのみ使用したプログラムで、各プログラム名の後ろに_RGCが付く、メルケプストラムに加えて回帰係数を使用したものもある。

（1）学習プログラム、単音節認識プログラム（孤立音声の認識用HMM）

孤立音声の認識用モデルとして混合連続出力分布型HMM（Continuous Mixture density Hidden Markov Model：CMHMM）の学習プログラムと認識プログラムを用意した。学習にはBaum-Welchの再推定アルゴリズムを用い、認識にはViterbiアルゴリズム(尤度の計算のみ)を用いている。入力されたすべてのカテゴリの音節HMMと、音声データをマッチングし、どのカテゴリのHMMで最も高い尤度を出すかを調べ、認識結果とする。

①学習プログラム（TR_CMHMM, TR_CMHMM_RGC）

Usage :

TR_CMHMM -i input-mel-file [-t hmm_table | -m init_hmm] -o out-hmm options

options : -s save_intermediate_file

-n min_iterations max_iterations

-h initial-hmm-output

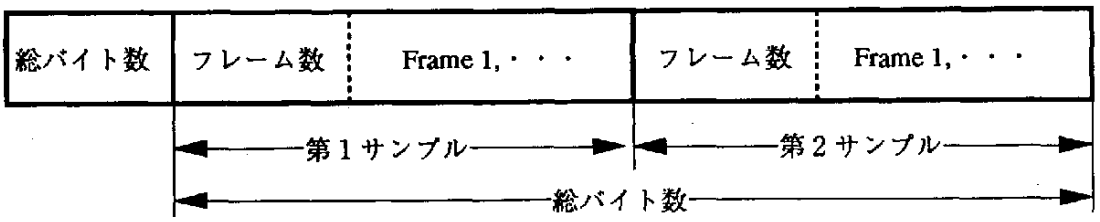
(-r random_ratio)

(-w rgc-width)

【説　明】

-i input-mel-file：

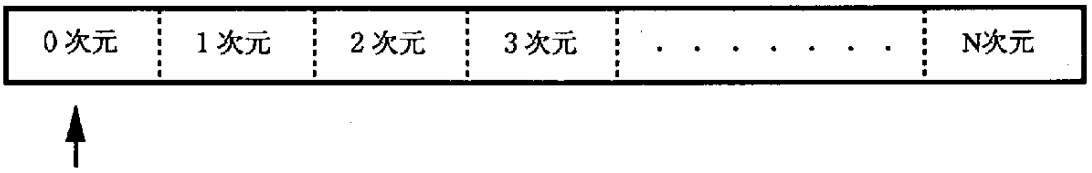
学習用データが格納されているファイル名。学習用、認識用、O(n)DPでは、ATRフォーマットを使用している。ATRフォーマットは、学習用データの各サンプルを次の書式で1つのファイルにマージしたものである。総バイト数はint型で、フレーム数はfloat型である。また各フレームはfloat型データが次元数だけシーケンシャルに並んだものである。



ATRフォーマット

各サンプルデータのフォーマットは、音声の特徴パラメータがN次元で表現されているとすると、パワー成分を付加した合計(N＋1)次元のフレーム系列がシーケンシャルに格納されて

いる。各フレームの各次元は float形式で格納されているものとする。



（0次元は読み込み時に無視される）

-t hmm_table または -m init_hmm：

hmm_table(ファイル名)あるいはinit_hmm(ファイル名)によりモデルの記述をおこなう。hmm_tableは初期モデルが存在しない場合のモデル構造の記述であり、-tスイッチhmm_tableを指定する。このhmm_tableは次のような書式となっている。なお、初期HMMはleft-to-right型で、状態に対して遷移するアーク全てが結びのモデルしか推定できない。

hmm_tableの内容例

```
10  4  1
 ¦  ¦  ¦
 ¦  ¦  +---混合連続出力分布の混合
 ¦  ¦        （この場合、混合数１つまり単一多次元正規分布）
 ¦  ¦       ＶＱを用いて混合分布の初期分布を推定する場合、
 ¦  ¦       ２分割繰り返しアルゴリズムを用いているため
 ¦  ¦       mixture = 2^m でなければならない。
 ¦  +-------left-to-right型HMMの出力分布数
 ¦           （状態数＝4 + 1
 ¦            この場合、5状態4出力分布型left-to-right型HMM）
 +----------音声特徴量ベクトルの次元数
             （この場合、１フレーム10次元の特徴量ベクトル）
```

init_hmmはhmm_tableを使った場合よりもより細かなモデルの記述をするとき、あるいは学習済みのモデルを初期モデルとして学習を進める場合に用いることができる。次に例を示してモデルパラメータについて説明する。

【CMHMMのモデルファイルのフォーマット】

10 ←特徴量ベクトルの次元数

4 ←出力分布の数

1 ←混合数

1.00 ←0番目の分布の分岐確率

1.81e+02 -2.87e+01 -1.49e+01 ・・・・ 4.11e-01 ← 平均ベクトル

1.64e+03 6.24e+01 -2.31e+02 ・・・・ 2.89e+00
6.24e+01 2.14e+03 -1.06e+03 ・・・・ 1.78e+02
-2.31e+02 -1.06e+03 1.52e+03 ・・・・ -3.65e+01
-4.13e+02 1.09e+03 -5.91e+02 ・・・・ 6.91e+01
2.51e+02 2.02e+02 -5.99e+02 ・・・・ -1.11e+02 ← 共分散行列
1.02e+02 3.88e+02 -5.06e+02 ・・・・ -2.45e+01
-1.59e+02 -2.05e+02 2.27e+02 ・・・・ -2.60e+01
1.65e+02 -2.22e+02 2.54e+02 ・・・・ 7.22e+01
-1.08e+02 5.26e+01 -1.52e+02 ・・・・ -1.42e+02
2.89e+00 1.78e+02 -3.65e+01 ・・・・ 1.72e+02

・

・（混合数だけ平均ベクトルと共分散行列が繰り返される）

・

・

5 ← 状態数

1 0 ← 初期状態の数とそのindex（この場合初期状態は1つで状態#0）

1 4 ← 最終状態の数とそのindex（この場合最終状態は1つで状態#4）

8 ← 遷移の総数

0 0 7.545061e-01 0
0 1 2.454939e-01 0 ← 遷移の記述（この場合、状態#0から状態#1への遷移は、
1 1 8.752607e-01 1 遷移確率は0.2454939、0番目の出力分布を用いる。
1 2 1.247393e-01 1
2 2 8.502138e-01 2

2 3 1.497862e-01 2

3 3 7.938289e-01 3

3 4 2.061711e-01 3

-o out-hmm :

学習済みのモデルのファイル名を指定する。

-s save_intermediate_file：学習途中の中間モデルを生成する。

学習の iterationで "save_intermediate_file".#なる中間モデルを生成する。

-r random_ratio：

ＶＱを用いないで初期分布を推定する場合（make_{mel,rgc}_nvqを使用）
混合連続モデルを用いる場合、初期モデルの平均ベクトルをばかす割合
（百分率で指定、デフォルトは１。大きすぎると学習に失敗する可能性が有る）。

-n min_iterations max_iterations：

学習回数の下限（min_iterations）と上限（max_iterations）を指定。
min_iterationは、最低でもその回数繰り返し、回数i > min_iteration
かつ、学習の尤度Lが、L(i-1) > L(i)となる、すなわち尤度が下がったら
終了する。以後尤度が上がり続ければ max_iterationsまで繰り返す。

-r rgc-width　　（make_rgcでコンパイル）

回帰係数の計算（内部でおこなう）する際の、前後何フレームを用いるかを
指定する。デフォルトは４。すなわち時刻tにおいて、t-4～t+4までの
フレームを用いて回帰係数を求める。

②単音節認識プログラ（DP_CMHMM, DP_CMHMM_RGC）

Usage : DP_CMHMM -h hmm-list -i input-file [-k syllable]

【説　明】

-h hmm-list：認識カテゴリの記述ファイル名の指定

hmm_listの例

3　　←カテゴリ数

model_ba ba ←１番目のモデルのファイル名がmodel_baで/ba/の音節

model_da da ←１番目のモデルのファイル名がmodel_daで/da/の音節

model_ga ga ←１番目のモデルのファイル名がmodel_gaで/ga/の音節

-i input-file：評価用データファイルの指定

-k syllable：現在与えた音声データのラベル

認識結果において、syllableのカテゴリが１番最初に出力される。

指定しない場合は、hmm_listに現れる順番に認識結果が出力される。

【認識例】

入力データ：/N/のカテゴリ、9個

Category size : 36

Observation filename : tem1m.mel.tmp [9 data sets (0 data rejected)]

Recognition rate

/　N/ = 44.44 %（　4 /　9）正解

/　e/ = 11.11 %（　1 /　9）誤り

/　ni/ = 44.44 %（　4 /　9）誤り

（２）O(n)DP法による連続音節の認識（音節HMMの連結法による音節認識）

オートマトン制御のない、すなわちすべての音節を出力する１状態のオートマトンを持つOne Pass DP法のHMM版連続音声認識アルゴリズム。ここでは学習済みの音節HMMを標準パターンとして用い、O(n)DP法により入力音声を最適な音節記号系列に変換する。

①認識プログラム（ONDP_CMHMM, ONDP_CMHMM_RGC）

Usage：ONDP_CMHMM -h hmm_list -im input-file

【説　明】

-h hmm_list：標準パターンとなるモデルファイルのリストを指定する（単音節認識用のものと同じ）。

```
110                 ←標準パターン数（この場合110）
model/hmm_N     N
model/hmm_Ti    Ti ←各標準パターンのファイル名と対応するシンボルの記述
model/hmm_a     a
      .
      .
      .
model/hmm_zyu   zyu
```

-im input-file：入力音声データのファイル名の指定

【認識例】

入力データ：メールを見たい

```
Category size : 110
Observation filename : onm1m.mel.tmp [1 data sets (0 data rejected)]
##### input : [1] length : 214 #####
  1 :   1 --->   39  39 /me/        音節番号 開始->終了フレーム フレーム 音節
  2 :  40 --->   65  26 /e/
  3 :  66 --->   80  15 /ru/
  4 :  81 --->  103  23 /to/
  5 : 104 --->  123  20 /to/
  6 : 124 --->  153  30 /mi/
  7 : 154 --->  184  31 /da/
  8 : 185 --->  214  30 /i/
##### log(P) = -8.199430e+03 #####  全体の尤度
```

【結果例】

```
sum.mix1m
#### Syllable Recognition by O(n)DP ####
speaker      SN    TI    HU    KO    MA    SE    ALL
-----------------------------------------------------
input       939   939   939   939   939   939   5634
-----------------------------------------------------
correct(%)  33.4  44.8  39.5  45.4  40.9  48.8   42.1
(num)      ( 314)( 421)( 371)( 426)( 384)( 458)(2374)
subst.(%)   52.3  47.0  49.4  46.2  50.0  42.7   47.9
(num)      ( 491)( 441)( 464)( 434)( 470)( 401)(2701)
insert(%)   22.4  27.0  28.5  25.7  30.2  38.7   28.8
(num)      ( 210)( 254)( 268)( 241)( 284)( 363)(1620)
delete(%)   14.3   8.2  11.1   8.4   9.1   8.5    9.9
(num)      ( 134)( 77)( 104)(  79)(  85)(  80)( 559)
-----------------------------------------------------
seg.rate(%) 63.4  64.8  60.4  65.9 60.7  52.8  61.3
```

（3）オートマトン制御One Pass DP法による連続単語認識（One Pass DP：OPDP）

音節HMMを連結した単語HMMをタスクに必要な数だけ作成し、その単語HMMとオートマトンより連続音声中の単語を認識する。

①認識プログラム（OPDP_CMHMM, OPDP_CMHMM_RGC）

Usage:

OPDP_CMHMM -h HMM-list -i input-file -d dictionary -s state_info
[-w rgc-width]

単語HMMリスト：HMM-list

flist_all_rob　　ロボット用

flist_all_num　　数字認識用

辞書ファイル：dictionary

word_dic_rob　　ロボット用

word_dic_num　　数字認識用

オートマトン情報ファイル：state_info

state_info_rob

state_info_rob1

state_info_num1

state_info_num2

state_info_num3

state_info_num4

辞書およびオートマトン情報ファイルのデータ構造：

辞書のデータ構造は単語シンボル、単語名、音節の3つ部分から構成されている。オートマトン情報ファイルは、まずオートマトンにもとづいて終了状態、そして各状態の番号と行き先の状態番号、およびその間のすべての単語シンボルの情報を記述している。

第４章　ツールの操作概要

第4章　ツールの操作概要

4.1　はじめに

本ツールは、AM79C30Aディジタル・サブスクライバ・コントローラチップを内蔵したSun SPARC Station 2、1等のワークステーション、およびdbriを搭載したSun SPARC Station 10用の音声処理支援ソフトウェアである。

具体的には、Sunのデモプログラムx_soundtoolと同様の音声の録音・再生、ファイル保存・ファイル読み込みをMotif環境上で実現し、さらに音声研究用の機能として、音声波形の拡大表示、スペクトル分析、ホルマントトラッキング・ラベリング編集、音声波形データの編集、および音声認識研究に使用するHMMを中心とした基本的なライブラリを付加したものである。

なお、扱う音声データのフォーマットはSunのAUDIOファイル（8KHz、8bit u,_law）に加え、研究用連続音声データベース（CD-ROM）などで採用されているIEEE形式（16bitヘッダーなしのベタファイル）である。この場合のサンプリング周波数は8KHz、10KHz、12KHz、16KHzの4種類である。

以下に、波形の拡大表示、スペクトル分析、ホルマントトラッキング・ラベリング編集、波形データの編集機能の操作概要について説明する。

4.2　機能概要

（1）録　音

Sunワークステーションのオーディオジャックに接続したマイク(またはライン信号)から音声信号をサンプリング周波数8KHz、8bit u,_law圧縮で取り込む。

スライダーによるレベル調整、録音の一時停止と再開(ポーズ)、オシロスコープ表示による信号のモニター、およびカウンタによる録音時間のモニター機能がある。

（2）再　生

Sunワークステーション内蔵のスピーカーまたはオーディオジャックに接続したヘッドホンから指定した範囲の音声信号を出力(再生)する。

スライダーによるレベル調整、再生の一時停止と再開(ポーズ)、ループ再生、オシロスコープ表示による信号のモニター、およびカウンタによる再生時間のモニター機能がある。

※ Sun の音声デバイスについては オンラインマニュアルの AUDIO(4)を参照。

※ 録音・再生のチャンネルはモノラルのみ。

（3）ファイル操作

Sun AUDIOファイル、IEEE 形式音声ファイルの読み書き、およびラベルファイル（テキスト形式）の読み書きをおこなう。

音声ファイルについては既存ファイルの後ろに追加して保存するアペンド機能もある。Sun AUDIOファイルは8KHz、8bit u_law 圧縮データで、IEEE 形式ファイルはサンプリング周波数が8、10、12、16KHzの16bitベタファイルである。

ラベルデータの時間の単位は1ポイントあたり 1.25、2.5msecの2種類である。詳細は4.5節「ラベルファイルデータ形式」を参照。

それぞれのファイルを修飾子を含めて個別に指定して読み書きするモードの他に、同一パス（ディレクトリ）上で、IEEE 形式音声ファイルとラベルファイルをそれぞれ ".ad"、".lb" のデフォルト修飾子を自動的に付けてまとめて扱うモードがある。また、波形の範囲の指定により全領域の保存と一部のみの保存が可能である。

（4）波形編集

指定した領域のコピー、削除、およびコピーまたは削除したデータ（カットバッファー上のデータ）をペーストする機能である。

（5）拡大波形表示

指定した領域の波形を拡大表示し、ラベル編集、スペクトログラム分析に用いる。なお、この機能とは別に、録音・再生時以外に波形表示上をマウス移動させると、マウスポインタ近傍の波形がオシロスコープに拡大表示される。

（6）ラベル編集

拡大波形表示上でラベル境界の挿入・削除、異なる層間でのコピー、量子化、およびラベル記号の編集をおこなう。ラベル境界の編集にはボックス範囲指定機能があり、ボックス内の一括削除や第1層から第2層へのコピーが可能である。また、ラベル記号の編集は、キー入力の他にマウス操作でメニューの中からラベル記号を選択することでも可能である。

（7）スペクトル分析

拡大波形表示上の波形全体、または、ボックス範囲指定した領域についてスペクトログラム分析をおこなう。表示は白黒／カラーのいずれかを選択し、バイアスで全体の濃淡を調整する。得られたスペクトログラム上にはホルマント位置の軌跡も表示される。

（8）コマンド

UNIXのコマンドの実行を容易にするヒストリー機能付きのサブウィンドウを表示する。画面のハードコピー（screendump）やディレクトリの内容のリスト出力（ls）などをツール上から実行できる。

4.3 画　面

voicetoolの画面は、メイン画面（図4.3-1）とサブ画面（図4.3-2）の2つから構成されている。ツール起動時にはメイン画面のみが表示され、音声データを録音またはファイル読み込み操作により取り込んだ後、波形画面上のポップアップメニューの中から拡大波形表示を選択することでサブ画面が表示される。なお、サブ画面が表示されているときに録音を開始すると、サブ画面は自動的に消える。

※ポップアップメニューは波形画面上にマウスポインタを移動させ、マウス右ボタンをクリックすることで表示される。なお、波形の表示がない場合と録音中、およびスペクトル分析中はポップアップメニューは表示されない。

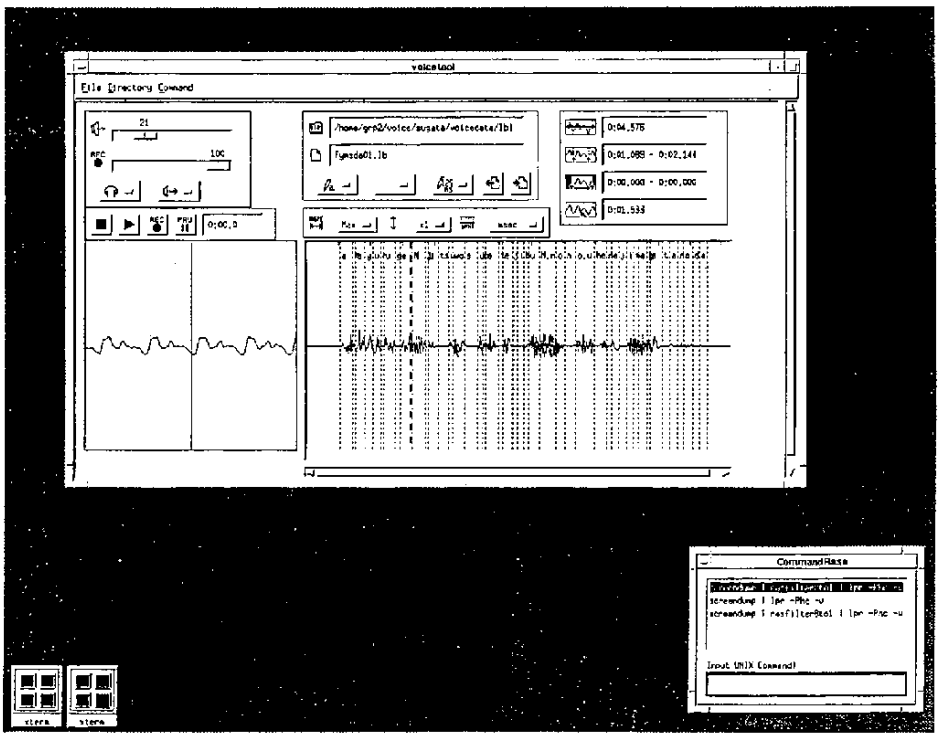


図4.3-1　メイン画面（音声の録音・再生操作、ファイル操作、波形編集）

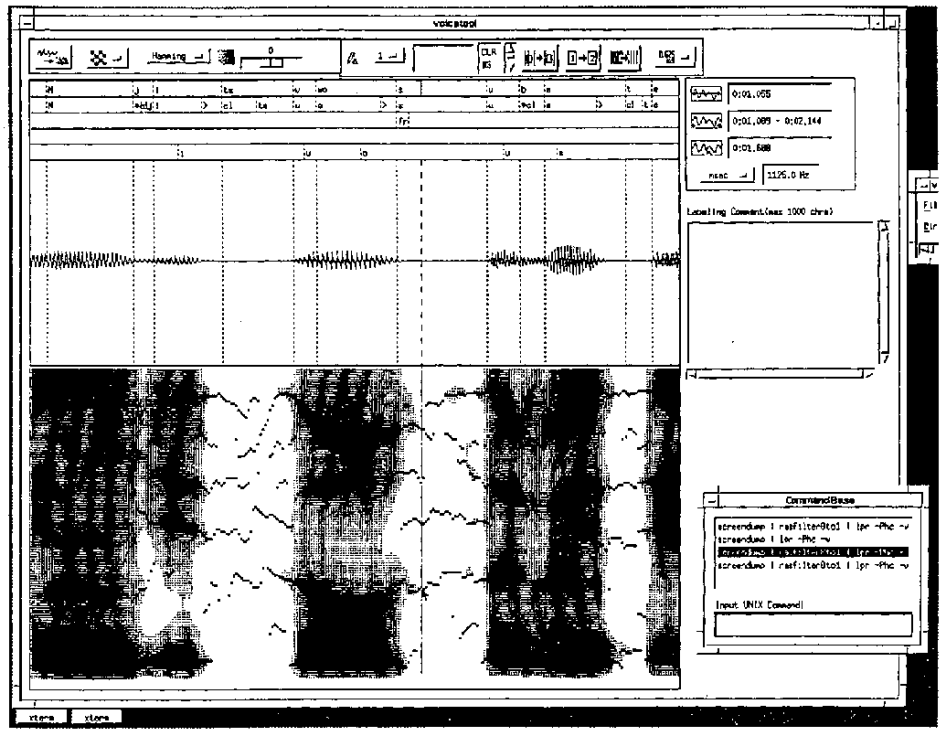


図４.３-２　サブ画面（ラベル編集、スペクトログラム分析）

４.４　操作説明と仕様

４.４.１　メインウィンドウ

（１）メインウィンドウ上のマウス操作

波形表示画面上のマウス操作は次のとおりである。

・マウス左・中ボタンクリック：波形選択領域の設定

・マウス右ボタンクリック　　：ポップアップメニュー表示

波形選択領域の設定はマウスクリックによる設定方法以外に次の方法がある。

A.ラベル境界に合わせるモード

このモードの切り替えはトグルスイッチになっている。

第１層目のラベルデータが存在する場合にこのモードを指定すると、マウスクリック位置の近傍のラベル境界位置に波形選択位置を合わせる。

※図４.４-１ Bの切り替えで第１層目のラベルデータを波形表示画面上に表示したり、消したりする。

C.開始点を先頭に移動

選択領域の開始点（左端）をデータの先頭にジャンプさせる。

D.開始点を最後に移動

選択領域の開始点（左端）をデータの最後にジャンプさせる。同時に終点（右端）もデータの最後にジャンプし、選択領域の幅は0になる。

波形編集のペーストでデータの最後にカットバッファー内容をペーストする場合に用いる。

E.終点を最後に移動

選択領域の終点（右端）をデータの最後にジャンプさせる。C（開始点を先頭に移動）との組み合わせで全領域の選択ができる。

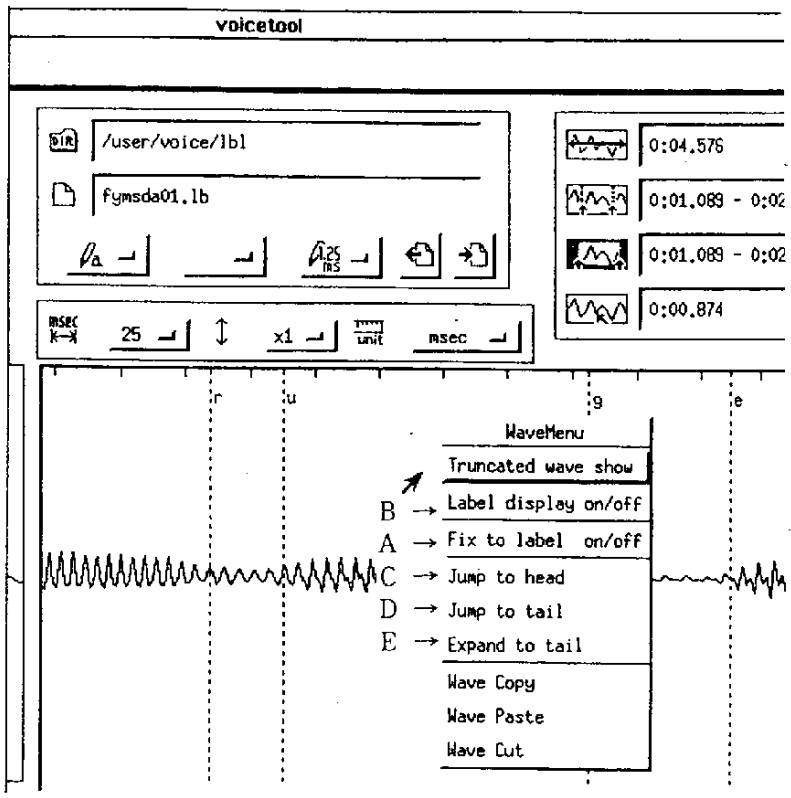


図4.4-1　メインウィンドウ上のマウス操作

（2）録音・再生

サンプリング周波数：8KHz

量子化ビット数　　：8bit u ,_law圧縮（精度12bit）

入出力モニター　：オシロスコープ、カウンター

レベル調整　：録音・再生のレベルをスライダーで設定

信号再生領域　：マウス指定

ループ再生　：オン・オフ切り替え

ポーズ機能

①ワーク領域上のボタン操作

A. 再生ボリューム

B. 録音ボリューム

C. 内蔵スピーカー/外部機器への出力 切り替え

他のオーディオ関係のアプリケーションと連動する。

D. ループ再生ON/OFF切り替え

ループ再生OFFで再生すると、データの最後で自動的に再生は停止する。

E. 録音・再生カウンタ

表示単位を（後述するボタン操作で）変更できる。

②オーディオ操作ボタン

操　作	前の状態	実行内容
F.停　止	停止中	な　し
	再生中	再生の停止
	録音中	録音の停止
	ポーズ中	停止（ポーズも解除）
G.再　生	停止中	再生開始。 先頭から再生
	再生中	な　し
	録音中	録音の停止後、直ちに先頭から再生
	再生ポーズ中	ポーズの解除。 ポーズ前の位置から再生
	録音ポーズ中	ポーズの解除。 先頭から再生
H.録　音	停止中	録音開始。 先頭から録音
	再生中	再生の停止後、直ちに先頭から録音
	録音中	な　し
	再生ポーズ中	ポーズの解除。先頭から録音
	録音ポーズ中	ポーズの解除。 ポーズ前の位置から録音

I.ポーズ	停止中	な し
	再生中	再生のポーズ
	録音中	録音のポーズ
	ポーズ中	ポーズの解除。ポーズ前の状態を続行

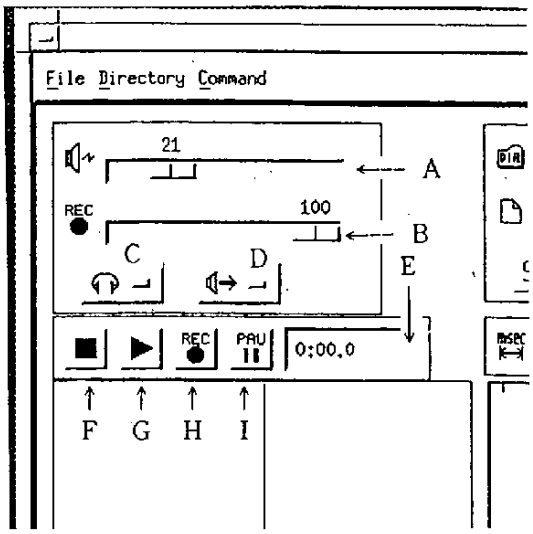


図4.4-2　録音・再生操作

（3）ファイル操作

①音声データの読み込み・保存、既存ファイルへのアペンド

フォーマット	Sun AUDIO形式	IEEE形式
ヘッダ	あり	なし
サンプリング周波数	8 KHz固定	8、10、12、16KHz
量子化ビット数	8 bit u ,_law圧縮	16bit（上位、下位の順）
推奨修飾子(※)	.au	.ad

※ファイル名の付け方は任意とするが、音声データ（IEEE形式のみ）とラベルデータを合わせて読み込み、保存するモードでは修飾子に".ad" が付いたファイルをIEEE形式の音声データとみなす。

②ラベルデータの読み込み・保存

・形　式　　：ATR提唱の形式（6層からなる階層テキスト形式）

・データ単位：1ポイントあたり 1.25 または 2.5msec

小数第1位まで有効のため最小単位はさらに10分の1

※ラベルデータの詳細は、4.5節「ラベルファイルデータ形式」で述べる。

A.ディレクトリ名の指定

B.ファイル名の指定

C.ファイル形式の選択

D.IEEE形式音声データのサンプリング周波数

E.ラベルデータ単位の選択

F.読み込み実行ボタン

G.保存実行ボタン

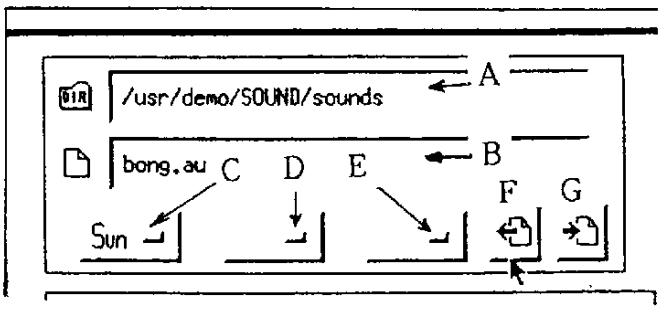


図4.4-3　音声データ／ラベルデータの読み込み・保存

※メニューバー上のFileメニューでラベルデータの読み込み・保存、音声データの読み込み・保存、アペンドの実行が可能。なおラベルデータのアペンドはできない。

③音声データ（IEEE形式のみ）とラベルデータの同時読み込み・保存

指定したファイル名にデフォルトの修飾子（音声データは".ad"、ラベルデータは".lb"）を自動的に付けてデータの読み出しと保存をおこなう。

※異なるディレクトリ上のファイルをまとめて扱うには上記の個別のファイル読み込み・保

存機能を使う。

※以下の例では、音声データ/usr/test/test.adとラベルデータ/usr/test/test.lbが保存される。

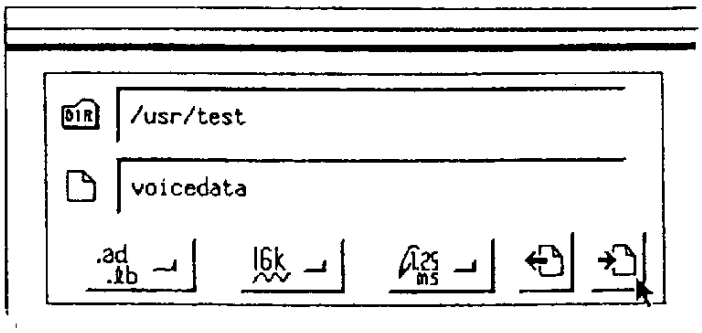


図4.4-4　音声データとラベルデータの同時読み込み・保存

④デフォルトディレクトリ名の保存・読み込み機能（メニューバー上のDirectoryメニュー）

2種類のディレクトリ名を保存し、読み出すことが可能である。また、カレントディレクトリ名をいつでも呼び出せる。

（4）波形編集・拡大波形表示

波形表示　　：時間軸（図4.4-5 A）と強度（図4.4-5 B）の倍率設定可能

拡大波形表示：指定領域の拡大表示（図4.4-5 C）

編集項目　　：指定領域のコピー（図4.4-5 D）・削除（図4.4-5 E）、カットバッファー内容のペースト（図4.4-5 F）

編集領域　　：マウス指定。ジャンプ機能（図4.4-5 G）あり。

※ペーストされる場所は、波形選択領域の先頭（開始点）である。

※ラベルデータが存在する場合は、ラベルデータも同時に編集される。

※編集機能の実行は波形表示画面上でマウス右ボタンをクリックすることで表示されるポップアップメニューの中から機能を選択しておこなう。

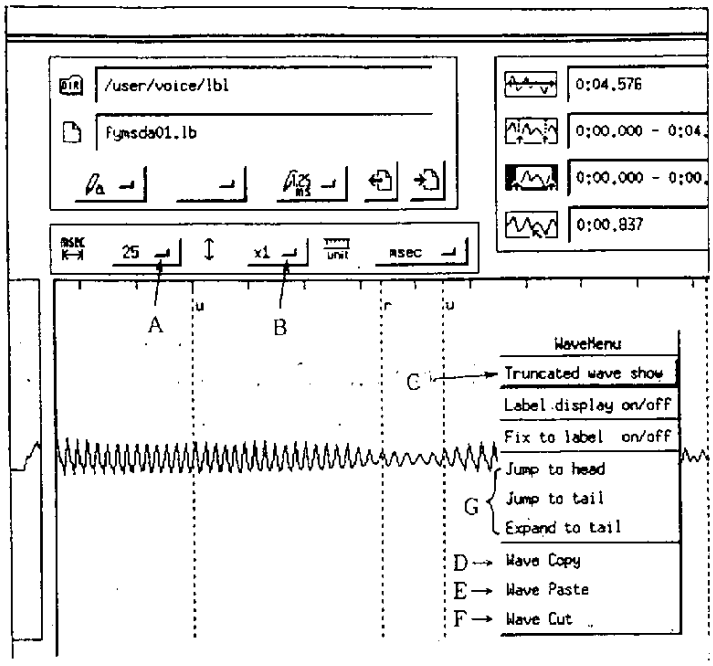


図４．４-５　波形編集・拡大波形表示

（５）オシロスコープ表示

録音・再生時：録音・再生時の波形を疑似的にリアルタイム表示

停止時　　：波形上にマウスがある場合、マウス位置の近傍の拡大表示

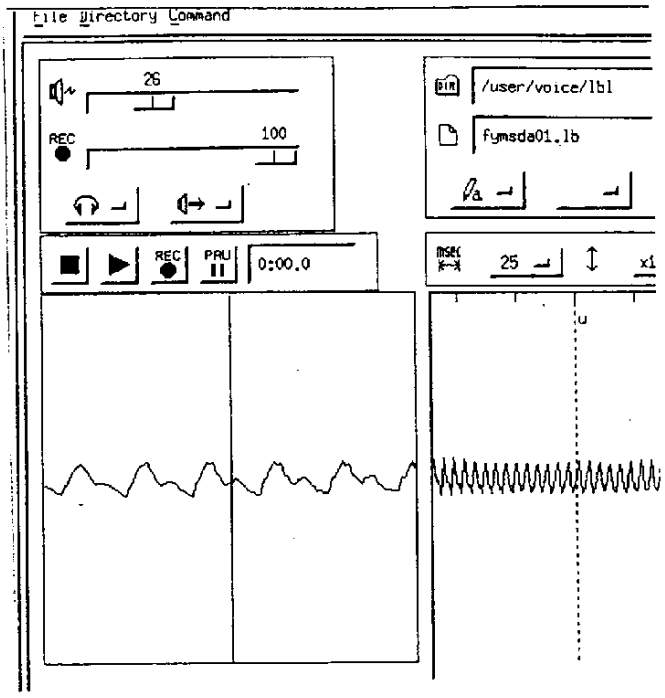


図４．４-６　オシロスコープ表示

（6）各種カウンタ表示

A.カウンタ単位設定：各種カウンタの表示データ単位を設定。msec では表示の最小単位が msec、/8 KHz では 1 ポイントあたり 1/8KHz、すなわち0.125msec を表わす。

B.録音/再生カウンタ 録音・再生位置（時間軸上位置）を表示する。

C.データ総量　　：データ量を表示する。

D.選択領域　　：データ指定領域の開始点と終点を表示する。

E.拡大波形カウンタ：拡大波形表示領域の開始点と終点を表示する。

F.マウス位置　　：波形表示上のマウス位置を表示する。

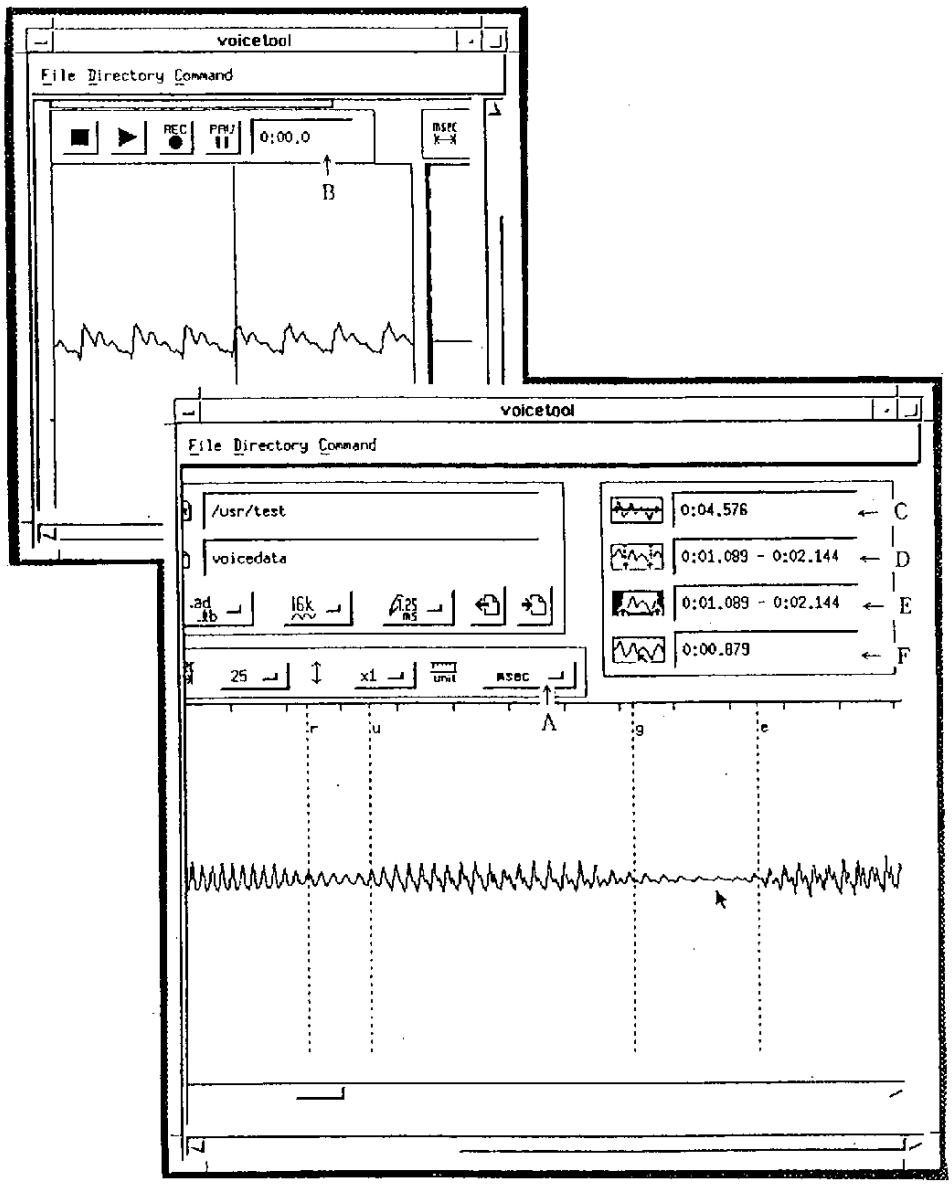


図４.４-７　各種カウンタ表示

（7）コマンド実行機能

UNIXのコマンドの実行をおこなうためのヒストリ機能付きのサブウィンドウを表示する。画面のハードコピー（screendump）やディレクトリの内容のリスト出力（ls）などをツール上から実行できる。

サブウィンドウ（コマンド・ダイアログ）は、コマンドヒストリエリア（図4.4-8 A）とコマンドラインエリア（図4.4-8 B）から構成され、コマンドラインエリアにコマンドをタイプするか、マウスによりコマンドヒストリエリアのエントリをクリックしたり矢印キーを使ってエントリを選択することでコマンドラインエリアにコマンドを表示させる。そしてリターンキーによりそのコマンドを入力する。入力されたコマンドはコマンドヒストリエリアに新たなエントリとして追加される。

また、コマンドの入力はコマンドヒストリエリアのエントリをダブルクリックしても可能である。

※ヒストリリストに保持可能な要素は最大10個までである。

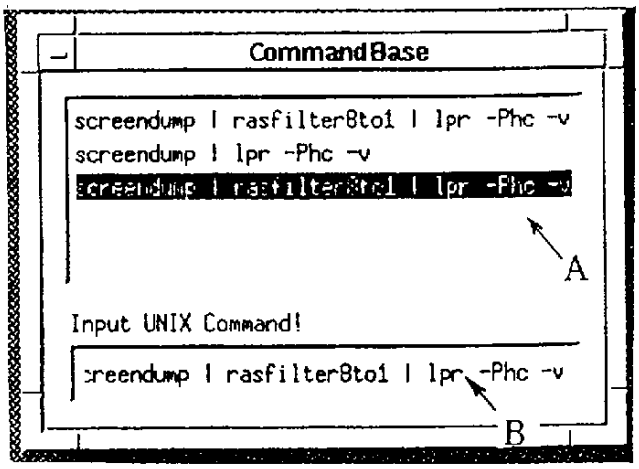


図4.4-8　UNIXコマンドの実行

4.4.2　サブウィンドウ

（1）サブウィンドウ上のマウス操作

各機能を実行する前に、まず拡大波形表示画面上でマウス右ボタンをクリックすることで表示されるポップアップメニューの中からマウスモードを選択する。

A.編集領域（Box領域）設定モード

B.ラベル境界編集モード

C.ラベル記号編集モード

D.音声再生範囲編集モード

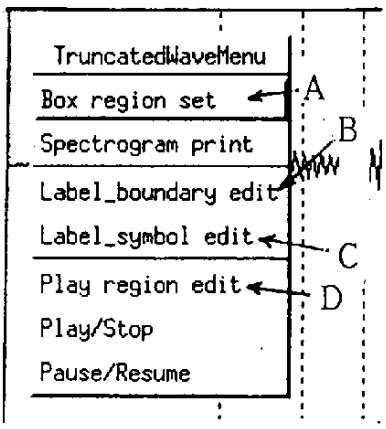


図４．４-９　サブウィンドウ上のマウス操作

また、以下の項目がウィンドウ上のボタン操作以外にもポップアップメニューの中から選択することで実行可能である。

A.スペクトル分析の実行

B.音声再生の開始と停止

C.音声再生のポーズと解除

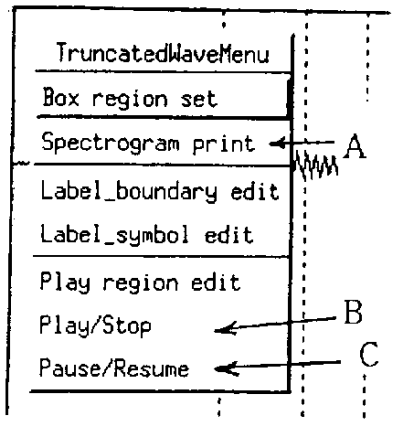


図４．４-１０

（2）ラベル編集

A.編集ラベル層切替：第1層データから第5層までの編集を切り替えておこなう。

B.編集ラベル層表示：現在の編集ラベル層を黄色（カラーモニターの場合）で表示する。

C.コメント層編集　：コメント層（第6層データ）を拡大波形表示の右側にあるエディタ画面内で編集する。

※コメント層の編集は最大1000文字までである（日本語非対応）。

D.ラベル境界編集　：マウス左ボタンクリック；

クリック位置に境界を挿入。このとき他の層の既存ラベル境界近傍をクリックすることでその内容（ラベル記号）と境界位置をコピーする。

マウス中ボタンクリック；

クリック位置近傍の境界を削除。十分離れた位置の境界は削除されない。

※上記操作の前にあらかじめマウスモードをラベル境界編集モードに設定する。

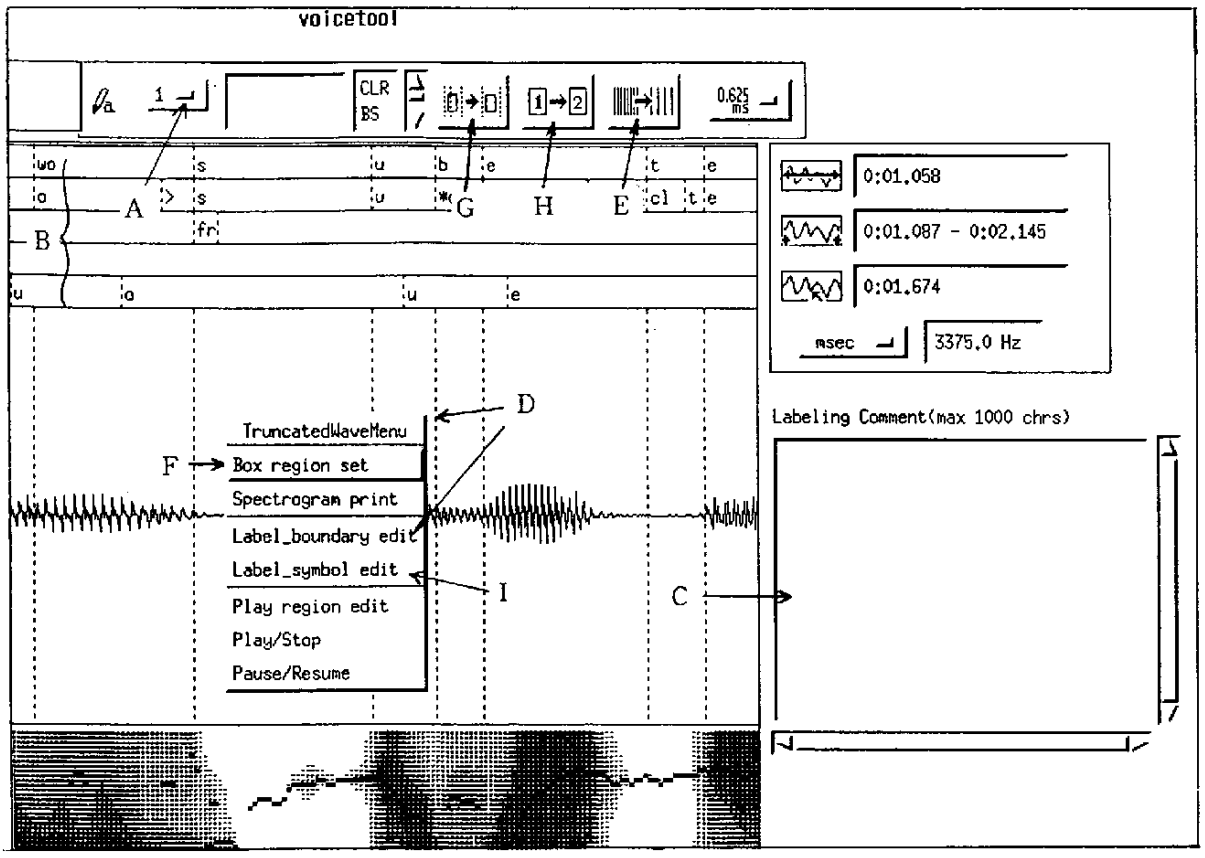


図4.4-11　ラベル編集

E.量子化　　：ラベル境界を拡大表示している全範囲の編集層についてのみ最小単位に合わせ込む。最小単位は、2.5msec、1.25msec、0.625msecのいずれかを選択できる。

なお、量子化をおこなわずにファイル保存した場合でも最小単位0.125msec（または0.25msec）への量子化が自動的におこなわれる。

F.ボックス領域編集　：マウスによる編集領域を指定した編集モード。指定幅が0の場合は全体が対象になる。

※ボックス領域指定による編集機能は、次のとおりである。

G.ラベル境界の削除　：

H.第1層データの第2層へのコピー（元々第2層にあるデータにマージ）

I.ラベル記号の編集　：キー入力またはメニューの中からラベル記号を選択する。

※この操作の前にあらかじめマウスモードをラベル記号編集モードに設定する。

（3）スペクトログラム分析

分析方法　　：線形予測（LPC）係数のFFTによるスペクトル包絡算出

表示白黒8諧調またはカラー8色による強度濃淡表示（図4.4-12B､E）

※バイアスで全体の濃淡を調整する（図4.4-12 C）。

ピーク（ホルマント）位置の軌跡表示する（図4.4-12 F）。

分析範囲　　：ボックス範囲指定した領域、または、拡大波形表示上の波形全体（領域幅が0の場合）

窓関数　　：ハミング窓、ハニング窓、矩形窓、三角窓(バートレット窓)、ブラックマン窓のいずれかを選択（図4.4-12 D）

フレーム長　　：32msec（8KHzで256ポイント）固定

周波数分解能　：31Hz固定

フレーム周期　：最小で 0.375msec（8KHzの3ポイント分）

※拡大波形表示倍率で変わる。

最大表示周波数：4KHz固定

分析次数　　：16

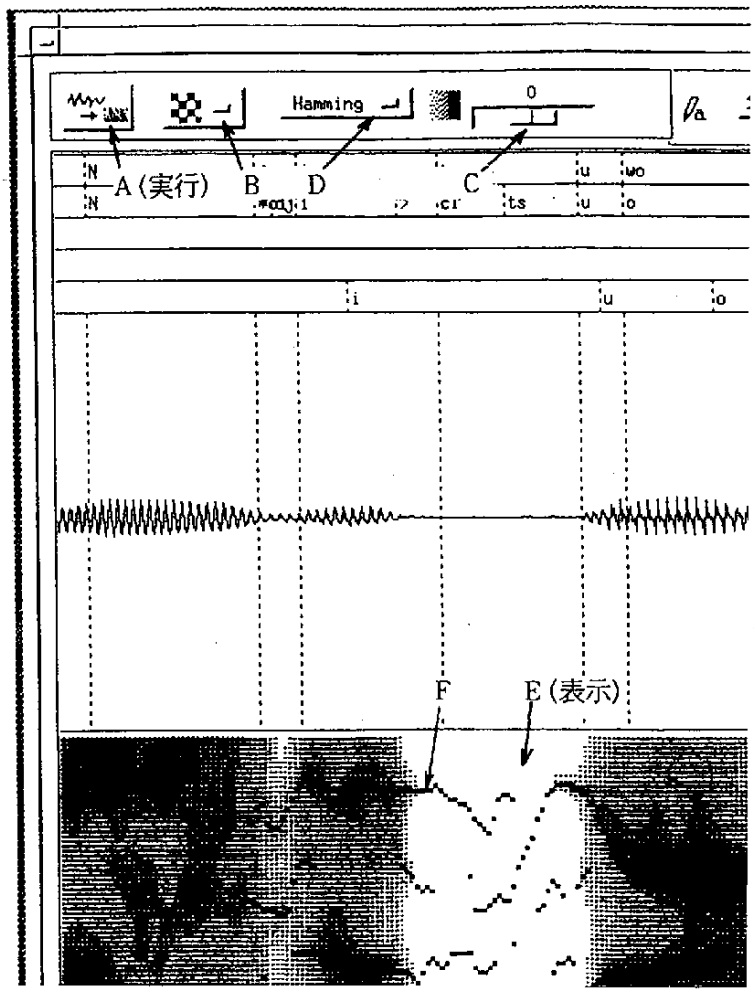


図４.４-１２　スペクトラム分析

（４）各種カウンタ表示

A.カウンタ単位設定：各種カウンタの表示データ単位を設定。msec では表示の最小単位が msec、/8 KHz では１ポイントあたり 1/8KHz、すなわち0.125msec を表す。

B.表示データ総量　：拡大波形表示の総量を表示する。

C.表示データ位置　：拡大波形表示の開始点と終点を表示する。

D.マウス位置　　：拡大波形表示上のマウス位置を表示する。

E.周波数表示　　：スペクトログラム結果表示上のマウスＹ方向位置(周波数)を表示する。
※表示単位は[Hz]固定である。

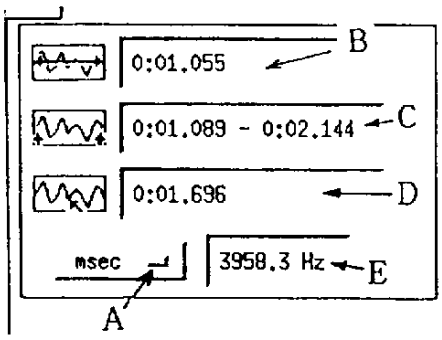


図４．４-１３　各種カウンタ表示

４．５　ラベルファイルデータ形式

音韻ラベルは基本的にはATRの提唱する形式に準じており、以下に示すように階層的に構成されている。文字コードはEUCである。

音声記号層	開始時間　区切り記号　ラベル記号　区切り記号　終了時間　CR ⋮
層区切り行	# CR
音声事象層	開始時間　区切り記号　ラベル記号　区切り記号　終了時間　CR ⋮
層区切り行	# CR
異音化層	開始時間　区切り記号　ラベル記号　区切り記号　終了時間　CR ⋮
層区切り行	# CR
融合事象層	開始時間　区切り記号　ラベル記号　区切り記号　終了時間　CR ⋮
層区切り行	# CR
母音中心層	中心時間　区切り記号　母音名　CR ⋮
層区切り行	# CR
コメント層	最大100文字までのテキストファイル形式（日本語非対応）

※「区切り記号」とは、任意の数のタブコードまたはスペースコードのことである。

※「CR」とは、改行コードである。すなわち１行分のデータで１組のデータが構成される。

4.6　ラベルファイルデータ詳細説明

（1）時間データ

「時間」データは数字と小数点から構成され、先頭が小数点の場合は「0.」と見なす。

※数字以外の文字や符号（マイナス）は使わないこと。

各層の中のデータの順番は開始時間（中心時間）順であらかじめソートされているものとする。

※複数のデータの範囲が重ならないようにする。

※「終了時間」が「開始時間」と等しい、あるいは小さい場合はそのデータは無視する。各データの開始時間は、その前にデータが存在する場合、前のデータの終了時間と等しいか、あるいは大きいものとする。母音中心層の場合は中心時間が等しいデータは後優先で前のデータが無視される。

時間の単位は、ポイントあたり1.25msec、または2.5msecの浮動小数点形式（0.1はそれぞれ0.125msec、0.25msecを表わす）である。

※時間単位の指定がvoicetoolのラベルファイル読み込みまたは書き込み直前に可能である。

（2）ラベル記号データ

「ラベル記号」には「区切り記号」や数字、小数点のコードがもともと含まれない。これらを「ラベル記号」内に入れた場合、終了時間が正しく扱えなくなる。

※「ラベル記号」の先頭が数字または小数点の場合はラベルが存在しないものと見なす。たとえば、境界の時間（開始時間、終了時間）のみ決定し、まだラベル記号が決まらないときにはラベル記号部分が省略される。

（3）コメント層データ

最大100文字までのテキストファイル形式とし、日本語には対応していない。100文字を越えるデータをvoicetoolに読み込むと、越えた分は無視される。

※「#」（層区切り文字）、「CR」（改行コード）、タブコード、数字、符号なども使用可能である。

（4）層構成

各層の区切りには「層区切り行」を1行のみ挿入する。この行の先頭はかならず「#」の文字で始める。なお「#」の後ろに文字列がある場合、その文字列は無視される。

各層の順番は上記の順番通りであるものとする。ある層のデータが全く存在しない場合でも「層区切り行」をかならず挿入する。

B.　音声認識関連製品

表B-1　音声認識ボード一覧（国内）

会社名	商品名	話者	発声	発声単位	語彙	登録	価格	発売時期	備考
NEC	P-3	特定	離散	単語	20語×2set	2回		90.1	自動車電話専用
富士通	FM77-421	特定	離散	単語	64語	1回以上	49,800円	86.	CPUに8088を用いた1ボード構成
	FM60-421	特定	離散	単語	92語	1回以上	10万円	88.12	録音・再生機能(CVSD方式,126秒) PB認識機能(12種,最大64桁) FMRシリーズ用ボード
日立	試作	特定	離散	単語	512語				制御：68008,照合：HD62Z800
	試作		連続	単語	128語				総IC数20個,音声合成付
シャープ	試作	特定	離散	単語	最大6千語程度 仮名文字設定可能			未定	16bitCPU,DSP,マッチング用LSI IBM PC,AX用ボード ADPCMによる録音・再生機能
松下通信工業	音声認識コントロールユニット	不特定	離散	単語	8語				壁掛形アンプ,または,ポータブルCDアンプの操作(再生・停止など)を専用ワイヤレスマイクから音声で行う 改良型連続線形伸縮照合方式［CLMα法］
NTT	SRP	特定	離散	単語	512語			未定	SA-DPとSPLIT法を採用 B5判の1ボード構成 電話音声を対象 同社プチアンサ用
	試作	不特定	連続	単語	80語			未定	単語HMMを用いた ワードスポッティングを採用 電話音声を対象 VMEバスインタフェース
リコー	「聞学者」MDB-SWR-02	特定	離散	単語	255語	3回	89,000円	92.7	PC9801用音声認識ボード ツールソフト付（音声キーエミュレータ，スピーチライブラリ）
	ユニットボードSWR-UB-02	特定	離散	単語	120語	3回	20万円 (ｻﾝﾌﾟﾙ価格)	92.7	機器組み込み用汎用ボード 2入力ノイズリダクション機能付 不特定対応可能
日本IBM	IBM日本語音声認識合成アダプター/A	特定（話者適応可能）	離散	単語	1,000語	標準3回	27.5万円	91.10	IBM PS/55用ボード (DOS/V用ソフトウェアを含む) 日本語テキスト音声合成, 録音・再生，電話機能を含む (OS/2 版ソフトウェアは別売)

表B-2　音声認識ボード一覧（国外）　(1/2)

会社名	商品名	話者	発声	発声単位	語彙	登録	価格	発売時期	備考
Technologia Systems Ltd.	Voice Strokes	不特定	連続					88.3	1チップ構成 音声タイプライタ用 録音再生機能あり
Texas Instruments	Julie	特定	離散	単語	16語		$100 以下	88.10	発声もできる多機能人形 TMS320C17ベースの カスタムDSPを使用
	TI speach	特定	離散 連続	単語 句	1,000語		$995	85.	IBM PC用, TMS32010使用 音声合成付 電話I/Fはオプション
Voice Connection	Introvoice V	特定	離散		400語		$495		IBM PC用
Voice Control Systems	VCS1000	不特定	離散	単語	42語		$500 ～$1,295	87.2	三つの用途別に語彙が用意されている ホームとキュリティ用（7語） FA・OA用（42語）(電話用30語) スタンドアロン型
	VR160/VR40	不特定	離散	単語	100語				50,000件の認識装置に利用されている
Votan Inc.	Tele Center	特定	離散	単語			$8,000 ～$18,000	88.3	ボイスメッセージングシステム IBM PC/XT用
	VPC2150	特定	連続		約 700語		788,000円		IBM PC用 音声応答, 電話I/F付
Westinghouse Voice Systemes	SA100	特定	連続		大語彙			88.3	100dB騒音以下でも認識可, メニュー形式の開発ツールあり, スタンドアロン型
Audec Corp	SSB-1000	特定	離散	単語	144語		$250		
Cere Data Research	Audi-i	不特定					$995	88.5	IBM PC/XT/AT用の 音声分析ソフトウェア TI-Speach カードを使用
Convox Inc.	Voice Master Key	特定	離散	単語 句	16語 ×16語	2回以上	$130	88.9	IBM PC/XT/AT用
Dialogic Corp.	VR/10	不特定	離散		16語		$3,695	88.4	Dialog/40(電話I/F) 同41を使用した音声ダイヤル VSC社の技術供与
DSP Group	Spin	不特定	離散	単語	60語		$2,600 （スタンドアロン型は$3,000）	87.5	TMS32020使用 HMMによる認識 IBM PC用, 認識率97% 応答時間200msec

表B-2　音声認識ボード一覧（国外）　(2/2)

会社名	商品名	話者	発声	発声単位	語彙	登録	価格	発売時期	備考
Dragon Systems Inc.	Voice Scribe-400	特定	離散	単語	400語		$995	87.5	IBM PC用ボード
	VoiceScribe-1000	特定	離散	単語	1,000語		$1,195		IBM PC用ボード
IBM Corp.	IBM Speech Sever Series	特定	離散	単語	20,000語			92.12 (Client/sever用) 93.5 (Stand alone)	IBMワトソン研究所で開発したTangoraシステムの製品化 Client/Server環境で使用する場合 SeverはIBM RS/6000 ClientはIBM RS/6000 またはPS/2 Stand aloneで使用する場合は IBM RS/6000用のボード2枚とソフトウェアが必要
	IBM Continuous Speech Series	不特定		文章	最大 1,000語			92.12	IBM RS/6000又は IBM PS/2用のソフトウェア ボード(M-ACPA)も必要
International Voice Products	Vocalink Recognition	特定	離散	単語	100語 計 400語		$1,295		IBM PC用ボード
	Vocalink 4000	特定	連続	単語	100語 計 200語		$5,200		IBM PC用ボード 文法とシンタッタスを用いた認識
ITT Delence Communication	VRS1280	特定	連続 離散		約 2,000語	1回以上		88.9	IBM PC/AT, VME用 音声合成機付 cpu:63000, TMS32020
Scott Instruments Corp.				単語			$9,000	88.5	IBM PC用
	SIR MODEL20	不特定	離散	単語	75語 計 200語		$1,495		IBM PC用 TMS320C25使用 電話I/Fはオプション

表B-3　音声認識装置一覧（国内）　(1/2)

会社名	商品名	話者	発声	発声単位	語彙	登録	価格	発売時期	備考
NEC	SR-2000	不特定	離散	単語	16語	――		88.1	電話オンラインサービス用
	SR-2500	不特定	離散	単語	16語	――		84.10	電話オンラインサービス用 SR-2000の小型システム
	SR-3200	不特定	離散	単語	16語			90.	電話オンラインサービス用
	DP3000	特定 特定 特定	離散 離散 連続	単語 単音節 単語	450語 50語	1回 3回 12回	430万円	86.1	オプションでADPCM音声出力機能
	SR-150	特定 特定	離散 連続	単語 単語	240語 35語	1回	39.8万円		オプションでADPCM音声出力機能
富士通	F2371B	特定 特定	離散 連続	単語 数字	256語 10語	標準2回 標準3回	130万円	87.8	PARCOR音声出力機能
	F2361B	特定 特定	離散 連続	単語 単語	4000語 50語	1回以上 1回以上	40万円	90.1	ADPCM録音再生機能 オプションで規則合成機能追加と， 語彙を12000 語に拡張可
	F2355L	不特定	離散	単語	17語	――		83.2	電話オンラインサービス用 PARCOR音声出力機能 規則合成機能
日立	VI-50	特定 特定	離散 連続	単語 単語	120語 10語		1100万円	84.	5チャンネル同時認識 ADPCM音声出力機能
	HR-150	不特定	離散	単語	16語	――		80.1	電話オンラインサービス用
シャープ	WD-2900V	特定 特定	離散 離散	単音節 単語	111語 50語	1回	345万円	84.12	音声ワープロ 8万語の辞書
		特定話者文節認識 自立語 34000語，カタカナ5000				6～10分			音声ワープロ ADPCM音声出力機能
	パーソナル ワープロ 「音声書院」	特定	5~6音節 /秒	文節 単語 英数字 機能語		111音節 +189単語		91.10	収容辞書約10万語 ADPCMによる操作ガイダンス
	コードレス 留守番電話機 CJ-A50	特定	離散	単語	20語	2回	13.4万円	92.9	連続DPによるワードスポッティング CELPによる録音再生
沖電気	MF-300	不特定	離散	単語	13語	――			電話オンラインサービス用 ADPCM音声出力機能

表B-3　音声認識装置一覧（国内）　(2/2)

会社名	商品名	話者	発声	発声単位	語彙	登録	価格	発売時期	備考
東芝	Power Voice	不特定	離散	単語	16語	――			録音編集 規則合成機能
	不特定話者 認識装置（試作）	不特定	離散	単語	500語	――			92年7月 VOICE SYSTEM WORLD WIDE'92にて展示
松下電気産業	VC171	不特定	離散	単語	最大 3,000語				データ検索用，実験用 改良型連続線形 伸縮照合方式［ CLMα法 ］
	音声予約リモコン	不特定	離散	単語	58語 （6群）				家庭用VTRの番組予約をリモコンとの音声対話で行う 改良型連続線形 伸縮照合方式［ CLMα法 ］
三菱電気	M6951	特定	離散	単語	1,000語	1回		86.5	ADPCM音声出力機能付 オプションで話者適応化機能
	AVI (試作)	不特定	離散	単語	1,000語 ×100セット 仮名文字 設定可能	――		未定	電話オンラインサービス用 (電話音声Q&Aシステム) 音声出力はADPCM編集合成/ 規則合成

表B-4　音声認識装置一覧（国外）

会社名	商品名	話者	発声	発声単位	語彙	登録	価格	発売時期	備考
Kurzweil AI	KVS-3000	特定	離散	単語	1000語	3 回	$6,000	85.9	
	Voice Report	特定	離散		20,000語		$18,900		
International Voice Products	VocaLink SYS300	特定	離散	単語	200語		$1,655		
	VocaLink Series4000	特定	連続	単語	200語		$5,200		
	VocaLink E-Z DIAL I	特定	離散	単語	40語		$995		音声ダイヤラー
Voice Control Systems	Telerec	不特定	連続	単語	50語	――	$3,000		
	Verbex series 5000,6000,7000	特定	連続	単語	80語～10,000語		$5,600 ～ $9,600		
Verbex Voice Systems	Series 5000	特定	連続	単語	100語		$5,600 ～ $6,950	87.9	Text-to-Speach機能
Worlds of Wonder, Inc.	Julie	特定	離散	単語			$100 以下	87.	音声出力人形
Atriculate Systems	Voice Navigator	特定	離散		1,000語		$1,300		
Convox Inc.	Voice master key	特定	離散		64語		$150		
Dragon Systems Inc.	Dragon Dictate-30K	特定	離散		30,000語以上				開発ツールあり,IBM PC用
	Dragon Writer-1000	特定	離散		1,000語以上		$3,600		開発ツールあり,IBM PC用 multi Speacker 話者適応 騒音適応 IBM PC用 Beta-test version Dragon Writer-1000と同じ
	ITT VRS 1280/PC	特定	連続		2,000語		$9,000		
Speach Systems Inc.	Phonetic Engine	不特定	連続		10,000語～40,000語		$10,500 ～ $47,100		
Votan Inc.	Voice Card	不特定 特定	連続		300語		$3,500		
Voice Processing Corp.	Voice Pro	特定	連続		13語		$5,000		
	VPC1000/V-Pro-4	不特定 特定	連続		1,000語	20～50語	$5,000		

表B-5 音声認識LSI（国内）

会社名	商品名	話者	発声	発声単位	語彙	登録	価格	発売時期	備考
日本電気	μPD7764D	特定	離散	単語	180語				2段DP部のLSI
	μPD7764D	特定	連続	単語	40語				ﾌﾟﾛｸﾞﾗﾑで推敲340単語まで拡張可能 分析用にμPD7763がある
	MC-4760, μ PD7761D, μ	特定	離散	単語	128語				チップセット
シャープ	LR3687	不特定	離散	単語	8語		1000円以下	85.10	ｱﾅﾛｸﾞ･ﾓｼﾞｭｰﾙと2チップ構成
沖電気	MSM6250	不特定	離散	単語	8語			85.	1チップ認識LSI
東芝	TC8860F	特定	離散	単語	10語				1チップ認識LSI(LAM内蔵)
	TC8861F TC8862F TMP80C50AU	特定	離散	単語	78語				
	TC8861F +TC8864F-00	特定	離散	単語	50語				
	TC8861F +TC8865F-01	特定	離散	単語	20語				騒音環境用
松下電子工業	MN1263	特定	離散	単語	62語			85.	ｳｫﾙｼｭ･ｱﾀﾞﾏｰﾙ 変換とDPﾏｯﾁﾝｸﾞ部のLSI
	MN19041A-VSTZ	不特定	離散	単語	15語			91.4	汎用DSPに音声認識プログラムを内蔵 LSIは当社製品,音声予約VTR, 壁掛けアンプに内蔵されている 改良型連続線形 伸縮照合方式〔CLMα法〕
NTT通研	NLC007	特定	離散	単語	512語		35,000円	91.	SA-DPとSPLIT用距離演算

C.　音声合成関連製品

【音声合成関連製品】

本資料は、社団法人日本電子工業振興協会 音声入力方式専門委員会からご提供いただいたものです。

表C-1　録音再生LSI(1)

メーカ	東芝	東芝	東芝	東芝	東芝	東芝	東芝
品番名	T6668	TC8831F	TC8832F	TC8834N/F	TC8835N/F	TC8830AF	TC8833F
符号化方式	ADM	ADM	ADM				
ビットレート	8～32kbps	11～32kbps	11～32kbps	11～32kbps	16～22kbps	8～32kbps	11～32kbps
接続メモリ	DRAM	DRAM	DRAM	DRAM	DRAM	SRAM	SRAM, PS-RAM
最大アドレス	1Mbit	16Mbit	16Mbit	4Mbit	2Mbit	8Mbit	8Mbit
プロセス技術	CMOS	CMOS	CMOS				
マイクアンプ ローパスフィルタ	内蔵	内蔵	内蔵	-	-	内蔵	内蔵
A-Dコンバータ	-	-	-	-	-	-	-
D-Aコンバータ							
原発振周波数	655kHz	655kHz	655kHz	655kHz	800kHz	512kHz	640kHz
標本化周波数							
電源電圧	4.5～5.5V	4.5～5.5V	4.5～5.5V	4.5～5.5V	4.5～5.5V	4.5～6.0V	4.5～5.5V
消費電流(typ) 動作時 スタンバイ時	1.4mA	2mA 0.1μA	2mA 30μA	1mA	2mA 2mA	2mA 0.2μA	4mA 0.2μA
パッケージ	60MFP	60MFP	60MFP	42DIP/44MFP	28DIP/28SOP	67FP	80FP
備考							

表C-2　録音再生LSI(2)

メーカ	TI	TI	三菱電気	日立	日立	シャープ	シャープ
品番名	TMS3477	TMS3478	M37451MC	HD81801	HD81802	LR3990	LR3991
符号化方式	CVSD	CVSD	ADPCM	ADPCM	ADPCM	ADPCM	ADPCM
ビットレート	13～98kbps	13～98kbps	12～28kbps	24/32kbps	32kbps	18～24kbps	18～32kbps
接続メモリ	DRAM	DRAM	DRAM	-	-	DRAM	SRAM
最大アドレス	6Mビット	4Mビット	512kbit	-	-	4Mbit	4Mbit
プロセス技術	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS
マイクアンプ ローパスフィルタ	無	無	無	無	無	内蔵	内蔵
A-Dコンバータ	-	-	8bit	-	-	10bit	10bit
D-Aコンバータ	10ビット	10ビット	8bit	-	-	10bit	10bit
原発振周波数	280～380Kkllz	560～760Kkllz	12.5MHz	16MHz	16MHz	8MHz	8MHz
標本化周波数				8kHz	8～10kHz	8kHz	6/8kHz
電源電圧	4.5～5.5V	4.5～5.5V	4.5～5.5V	4.5～5.5V	4.5～5.5V	4.5～5.5V	4.5～5.5V
消費電流(typ) 動作時 スタンバイ時	2.2mA	3.3mA	40mA	100mA 10μA	100mA 10μA	5mA	-
パッケージ	28SDIP/SOP	28SDIP/SOP	64DIP 80QFP	40DIP	40DIP	64FP	80FP
備考				CODEC外付け、 音声区間検出付	CODEC外付け、 音声区間検出付	無音圧縮可、 メモリ管理機能	無音圧縮可

表C-3　録音再生LSI(3)

メーカ	松下電子工業	沖電気工業	沖電気工業	沖電気工業	沖電気工業	NEC
品番名	MN6410	MSM6588	MSM6388	MSM6258	MSM6310	μPD77C30C
符号化方式	ADM	3/4bitADPCM	4bitADPCM	3/4bitADPCM	8bitPCM	ADPCM
ビットレート	16.7～26.7kbps	12～64kbps	16～64kbps	12～62.4kbps	32～128kbps	24/32kbps
接続メモリ	SRAM	専用メモリ	専用メモリ	DRAM/SRAM	DRAM	外部制御
最大アドレス	1Mbit	4Mbit	4Mbit	16M(DRAM)/4M(SRAM)	512kbit	外部制御
プロセス技術	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS
マイクアンプ ローパスフィルタ	内蔵	内蔵 -40dB/Oct	内蔵 -40dB/Oct	無	ラインアンプ -24dB/Oct	無
A-Dコンバータ	-	12bit	12bit	8bit	8bit	-
D-Aコンバータ	-	12bit	12bit	10bit	8bit	-
原発振周波数	-	4.0～8.092MHz	1.5～4MHz	4～8MHz	4～8.2MHz	8.192MHz
標本化周波数		4～16kHz	4～16kHz	4～15.6kHz	4～16kHz	
電源電圧	4.5～5.5V	3.5～5.5V	3.5～5.5V	3.5～6V	4.5～5.5V	4.75～5.25V
消費電流(typ) 動作時 スタンバイ時	 5mA 10μA	 15mA 10μA	 10mA 10μA	 4mA 10μA	 10mA 10μA	
パッケージ	64FP	44QFP	44QFP	40DIP/44QFP/60QFP	44QFP	28DIP
備考	-					

表C-4　音声合成LSI(1)

メーカ	東芝	東芝	東芝	三洋電機	三洋電機	松下電子工業	松下電子工業
品番名	TC8801N/F	TC8804N/F	TC8802AF/BF	LC8100	LC8200	MN6401	MN6402
符号化方式	ADM	ADM	PARCOR	PARCOR	PARCOR	PARCOR	PARCOR
ビットレート	11～32kbps		1.7～9.8kbps	2.4, 4.8, 9.6kbps	2.4, 4.8kbps	1.2～5.5kbps	～2.75kbps
接続メモリ	256kbit内蔵	128kbit内蔵	64kbit内蔵	32kbit内蔵	外付ROM 512kbit	内蔵32kbit	内蔵8kbit
最大アドレス	-	-	8Mbit			-	-
プロセス技術	CMOS		CMOS			NMOS	CMOS
D－Aコンバータ			9bit	9bit	9bit	極性＋8bit	極性＋8bit
ローパスフィルタ	無		内蔵			無	無
原発振周波数	640kHz	640kHz	800kHz	400kHz		800kHz	800kHz
標本化周波数				9.1kHz	8kHz	10kHz	10kHz
電源電圧	2.4～5.5V	2.4～5.5V	3.9～5.7V	4.5～5.5/2.7～3.5V	4.5～5.5V	4.75～5.25V	3.5～6.0V
消費電流(typ) 動作時 スタンバイ時	 440μA 0.2μA	 440μA 0.2μA	 1.9mA 0.2μA	 2mA 10μA	 3.5mA 2mA	20mA	 3mA 10μA
パッケージ	28DIP	28DIP	44MFP	64QIP/24DIP/28SOP	40DIP/48QIP	28DPI	22DPI
備考							

表C-5　音声合成LSI(2)

メーカ	沖電気工業	沖電気工業	沖電気工業	沖電気工業
品番名	MSM6652/53/54/55/56	MSM6650	MSM6378A/79	MSM6585
符号化方式	4bitADPCM/8bitPCM	4bitADPCM/8bitPCM	4bitADPCM	4bitADPCM
ビットレート	16 ～ 256 kbps	16 ～ 256 kbps	16 ～ 128 kbps	16 ～ 128 kbps
接続メモリ	288k/544k/1M/1.5M/2Mbit内蔵	ROM	256k/512kbitOTP内蔵	－
最大アドレス	－	64Mbit	－	－
プロセス技術	CMOS	CMOS	CMOS	CMOS
D－Aコンバータ	12bit	12bit	12bit	12bit
ローパスフィルタ	内蔵(-40dB/oct)	内蔵(-40dB/oct)	内蔵(-24dB/oct)	内蔵(-40dB/oct)
原発進周波数	4.096MHz	4.096MHz	64k ～ 256kHz	640kHz
標本化周波数	4 ～ 32 kHz	4 ～ 32 kHz	4 ～ 32 kHz	4 ～ 32 kHz
電源電圧	2.7 ～ 5.5 V	2.7 ～ 5.5 V	2.7 ～ 5.5 V	2.7 ～ 5.5 V
消費電流(typ)				
動作時	15 mA	15 mA	20 mA	10 mA
スタンバイ時	10 μA	10 μA	10 μA	－
パッケージ	ﾁｯﾌﾟ・18DIP・24SOP	64QFP	16DIP・32SSOP/32SSOP	18DIP
備考	メロディー機能			

表C-6　音声合成LSI(3)

メーカ	シャープ	シャープ	シャープ	シャープ	シャープ
品番名	LR3681	LR3683	LR3985	LR3984	LR3987
符号化方式	波形素片	ADPCM	ADPCM	ADPCM	ADPCM
ビットレート	1.6 ～ 3.2 kbps	15 ～ 55 kbps	12 ～ 32 kbps	12 ～ 32 kbps	12 ～ 32 kbps
接続メモリ	DRAM 32kbit内蔵	FIFO内蔵	外付ROM	1Mbit内蔵	256kbit内蔵
最大アドレス	128kbit	制限なし	8Mbit	－	－
プロセス技術	CMOS	CMOS	CMOS	CMOS	CMOS
D－Aコンバータ	8bit	8bit	10bit	10bit	10bit
ローパスフィルタ	無	無	無	無	無
原発進周波数	4.1934MHz	4.1943MHz	2MHz	2MHz	2MHz
標本化周波数		7.2kHz	4 ～ 8 kHz	4 ～ 8 kHz	4 ～ 8 kHz
電源電圧	-3.3 ～ -2.7 V	4.5 ～ 5.5 V	2.7 ～ 5.5 V	2.7 ～ 5.5 V	2.7 ～ 5.5 V
消費電流(typ) 動作時 スタンバイ時	 1.5 mA 1 μA	 5 mA 3 μA	 4 mA 2 μA	 4 mA 2 μA	 4 mA 2 μA
パッケージ	48QFP	24DIP	60FP	24SOP	24SOP
備考		メロディ発生可能	メロディ発生可能	メロディ発生可能	メロディ発生可能

表C-7　音声合成LSI(4)

メーカ	TI	TI	TI	TI	TI	TI	TI	TI
品番名	TSP50C10	TSP50C11	TSP50C41	TSP50C42	TSP50C43	TSP50C44	TSP53C30	TMS50C20
符号化方式	LPC/PCM	LPC/PCM	LPC	LPC	LPC	LPC	LPC/PCM	LPC
ビットレート	1.3～8.6kbps	1.3～8.6kbps	1.3～8.6kbps	1.3～8.6kbps	1.3～8.6kbps	1.3～8.6kbps	1.3～8.6kbps	1.3～8.6kbps
接続メモリ	内蔵64kbit	内蔵128Kbit	内蔵64Kbit	内蔵64Kbit	内蔵128Kbit	内蔵128Kbit	-	EPROM
最大アドレス	-	-	-	-	-	-	-	-
プロセス技術	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS
D－Aコンバータ	10ビット	10ビット	8.5ビット	8.5ビット	8.5ビット	8.5ビット	8.5ビット	8.5ビット
ローパスフィルタ	無	無	無	無	無	無	無	無
原発振周波数	7.68/9.6MHz	7.68/9.6MHz	3.07/3.84MHz	3.07/3.84MHz	3.07/3.84MHz	3.07/3.84MHz	3.07/3.84MHz	3.07/3.84MHz
標本化周波数	8/10KHz	8/10KHz	8/10KHz	8/10KHz	8/10KHz	8/10KHz	8/10KHz	8/10KHz
電源電圧	4.0～6.5V	4.0～6.5V	4.0～6.0V	4.0～6.0V	4.0～6.0V	4.0～6.0V	4.0～6.0V	4.0～6.0V
消費電流(typ) 動作時 スタンバイ時	 5mA 10μA(max)	 5mA 10μA(max)	 1.5mA 10μA	 1.5mA 10μA	 1.5mA 10μA	 1.5mA 10μA	 1.5mA 10μA	 1.5mA 10μA
パッケージ	16ピンDIP/SSOP	16ピンDIP/SSOP	28ピンSDIP	40ピンSDIP	28ピンSDIP	40ピンSDIP	28ピンSDIP	40ピンSDIP
備考	8bitCPU内蔵	8bitCPU内蔵	8bitCPU内蔵 AMP内蔵	8bitCPU内蔵 AMP内蔵	8bitCPU内蔵 AMP内蔵	8bitCPU内蔵 AMP内蔵	ｽﾚｰﾌﾞﾓｰﾄﾞ動作	外付けEPROM

表C-8　音声合成LSI(5)

メーカ	NEC	NEC	NEC	NEC
品番名	μPD7755	μPD7756	μPD7757	μPD7759
符号化方式	ADPCM, PCM	ADPCM, PCM	ADPCM, PCM	ADPCM, PCM
ビットレート	16 ～ 32 kbps	16 ～ 32 kbps	16 ～ 32 kbps	16 ～ 32 kbps
接続メモリ	96 kビット内蔵	256 kビット内蔵	512 kビット内蔵	外付けDRAM
最大アドレス	－	－	－	1Mビット
プロセス技術	CMOS	CMOS	CMOS	CMOS
D－Aコンバータ	9ビット	9ビット	9ビット	9ビット
ローパスフィルタ	無	無	無	無
原発進周波数	640 kHz	640 kHz	640 kHz	640 kHz
標本化周波数	4, 5, 6, 8 kHz	4, 5, 6, 8 kHz	4, 5, 6, 8 kHz	4, 5, 6, 8 kHz
電源電圧	2.7 ～ 5.5 V	2.7 ～ 5.5 V	2.7 ～ 5.5 V	2.7 ～ 5.5 V
消費電流(typ) 動作時 スタンバイ時	 2 mA 20 μAMAX.	 2 mA 20 μAMAX.	 2 mA 20 μAMAX.	 10 mA 20 μAMAX.
パッケージ	18ピンDIP 24ピンSOP	18ピンDIP 24ピンSOP	18ピンDIP 24ピンSOP	40ピンDIP 52ピンQFP
備考				マイクロコンピュータ インタフェースあり 外部から音声データ 転送可能

表C-9　規則音声合成応用製品(1)

メーカ	沖電気工業	日本ＩＢＭ	リコー	KDD/ﾊﾄﾞｿﾝ/大井電気	ﾏｲｸﾛﾆｸｽ	明電舎
品番名	ＴＳ７５０Ｂ	IBM日本語音声認識・合成アダプター／Ａ	ＶＣ－１	TALK-J200	音次郎	FINE TALK
システム構成・形状	ユニット	IBM PS/55用ﾎﾞｰﾄﾞ+ｿﾌﾄ	ユニット	PC9801内蔵ﾎﾞｰﾄﾞ+ﾜｰﾌﾟﾛ・ｿﾌﾄ		ユニット
合成方式	波形重畳法	LSP-CVVC	ＬＳＰ	ﾎﾙﾏﾝﾄ	PARCOR	音響管
入力形態	漢字かな混じり文 韻律記号付かな文字列 JIS/ｼﾌﾄJIS/EUC	漢字かな混じり文	漢字かな混じり文*	漢字かな混じり文	JISｺｰﾄﾞ	漢字かな混じり文 韻律記号付かな文字列
発声スピード	10段階	10段階	9段階	7段階		7段階
発声ピッチ	8段階 男声／女声	男声	男声 9段階	男声／女声		男声／女声
搭載辞書	約６万語 ユーザ辞書約7500語	約 7万語	7万語*	約10万語 ﾎｽﾄ・ﾒﾓﾘ使用	約5万語 ﾎｽﾄ・ﾒﾓﾘ使用	約8.5万語
出力形態	内蔵スピーカ ヘッドホン 外部接続端子	外部接続端子	外部接続端子 ヘッドホン端子	内蔵スピーカ 外部接続端子		内蔵スピーカ イヤホン端子 ﾗｲﾝ出力
インタフェース	ＲＳ２３２Ｃ	マイクロチャンネル	RS232C	ﾊﾟﾗﾚﾙﾎﾟｰﾄ	ｾﾝﾄﾛﾆｸｽ	RS232C
価格		27.5万円（DOS/V用）	98,000円+128,000円*	15万円	12.8万円	19.8万円
備考		音声認識(1,000語),録音再生,電話機能を含む (OS/2版ｿﾌﾄｳｪｱは別売)	*:日本語処理部として言語工学研究所の校了を用いた場合。	卓上型TALK-J100もある（35万円）	ﾌﾛﾝﾄﾌﾟﾛｾｯｻ付	

表C-１０　規則音声合成応用製品(2)

メーカ	富士通	富士通	富士通	富士通
品番名	FMVS-101	F2361B+F2361B11	FM60-431	FMVS-101+FMTALK II
システム構成・形状	ユニット 60×200×160mm	ユニット内蔵ボード 130×358×230mm	パソコン内蔵ボード	ユニット 60×200×160mm
合成方式	PARCOR残差駆動	PARCOR残差駆動	PARCOR	PARCOR残差駆動
入力形態	漢字仮名混じり文*1	漢字仮名混じり文*1	漢字仮名混じり文*1	漢字仮名混じり文 *1
発声スピード	10段階	10段階	10段階	10段階
発声ピッチ	男声/女声各5段階	男声/女声各5段階	男声/女声各5段階	男声/女声各5段階
搭載辞書	約7万語 *1	約7万語 *1	約7万語 *1	単漢字辞書 *1
出力形態	内蔵スピーカ	内蔵スピーカ イヤホン	スピーカ用ジャック ヘッドホン端子・LINE-OUT	内部スピーカ イヤホン
インタフェース	セントロニクス	GP-IB *2	内部バス	セントロニクス
価格	7万円+2.8万円 *1	50万円+18万円 *1	3万円+2.8万円 *1	7万円+9万円*1
備考	*1:上位のパソコンで動作するソフトウェア日本語処理部として、Voice Managerを用いた場合	*1:上位のパソコンで動作するソフトウェア日本語処理部として、Voice Manager開発システムを用いた場合 *2:音声入出力装置用音声合成ボード	*1:上位のパソコンで動作するソフトウェア日本語処理部として、Voice Managerを用いた場合	*1:文字の読みあげ拡大表示ソフトウェア FMTALK II とVoice Manager ランタイムシステムを用いた場合(上位のパソコン上で動作するソフトウエア)

表C-１１　規則音声合成応用製品(3)

メーカ	NTT	NTT	NTT	NTT	三洋電機
品番名	しゃべりん坊 ﾀｲﾌﾟ1	しゃべりん坊 ﾀｲﾌﾟ2E	しゃべりん坊　ＨＧ	しゃべりん坊 ﾎﾟｰﾀﾌﾞﾙ	VSS-300
システム構成・形状	PC9801内蔵ボード	PC9801内蔵ボード	PC9801内蔵ボード	外部(50x100x170)	ﾕﾆｯﾄ
合成方式	LSP-CV-VC	LSP-CV-VC	LSP-COC	LSP-CV-VC	単音節規則合成
入力形態	漢字仮名交じり文 アクセント付き仮名文	漢字仮名交じり文 アクセント付き仮名文	漢字仮名交じり文 アクセント付き仮名文	漢字仮名交じり文 アクセント付き仮名文	ｶﾀｶﾅ列、数字 ｱﾙﾌｧﾍﾞｯﾄ
発声スピード	８段階	８段階	８段階	８段階	5段階
発声ピッチ	男声／女声	男声／女声	男声／女声	男声／女声	5段階
搭載辞書	約４．５万語、 RAM/ROM	約４．５万語、ROM	約６．５万語、 RAM/ROM	約４．５万語、 RAM/ROM	無
出力形態	ヘッドホン、 外部接続端子	ヘッドホン、 外部接続端子	ヘッドホン、 外部接続端子	ヘッドホン、 外部接続端子	内蔵ｽﾋﾟｰｶ 外部接続端子
インタフェース	パラレルポート	パラレルポート	パラレルポート	セントロニクス （プリンタ端子）	セントロニクス
価格	198000円	98000円	258000円	98000円	59800円
備考	PC9801対応	PC9801対応	PC9801対応	PC9801対応, 430g, 電池駆動可能	

表C-１２　規則音声合成試作装置(1)

メーカ	沖電気工業	キヤノン	シャープ	ＮＥＣ	東芝
品番名		－		規則合成装置(仮称)	試作機
システム 構成・形状	ボード(6×9cm)	ＮｅＸＴ用ｿﾌﾄｳｪｱ	ユニット	ユニット＋ホスト	J-3100(ISA)ボード
合成方式	波形重畳法	メルケプストラム	ホルマント	残差制御型CV-VC方式	ケプストラム-CV
入力形態	漢字かな混じり文 韻律記号付きかな文字列 JIS/ｼﾌﾄJIS/EUC	漢字かな交じり文	漢字かな混じり文	韻律記号付ﾛｰﾏ字 (ﾎｽﾄとのｲﾝﾀﾌｪｰｽ)	漢字仮名混じり文 ｱｸｾﾝﾄ付音声記号例
発声スピード	10段階	可変	８段階	11段階	16段階
発声ピッチ	8段階 男声／女声	男声	男声	11段階 女声	16段階 男声／女声
搭載辞書	約６万語 ユーザ辞書約7500語	約１０万語	約８万語	－	約10万語
出力形態	ヘッドホン	ＮｅＸＴスピーカ	内蔵スピーカ 外部接続端子	電話回線 (モジュラージャック)	外部接続端子
インタフェース	ＩＣカード	－	ＲＳ２３２Ｃ	ＲＳ２３２Ｃ	パラレルポート RS232C セントロニクス
備考	電池駆動 超小型テキスト・ 音声変換ボード			録音再生機能、ﾌｧｸｼﾐﾘ 応答装置接続機能、 ﾌﾟｯｼｭﾎﾝによる制御機能	SPARC station互換機用 文-音声変換ソフトもある

D.　入手可能な音声データ

【入手可能な音声データ】

(1)日本音響学会研究用連続音声データベース

((財)日本情報処理開発協会 AI・ファジー振興センター)

CD-ROM 16kHz標本化

Vol.1-3：ATR 音素バランス 503文 延べ 9600文、男性30名,女性34名

Vol.4-6：案内タスク文 延べ 11500文、男性18名,女性18名

Vol.6 ：案内タスク文+模擬対話テキスト

(2)電子協騒音データベース

((社)日本電子工業振興協会)

各種環境騒音、 DAT 18巻

(3)電子協日本語共通音声データ

((社)日本電子工業振興協会)

単音節、地名、数字、4桁数字、制御語、計 323項目

男女各75名、4回発声、DAT 76巻

参考文献

[1] 板橋「騒音データベースと日本語共通音声データ DAT版」
日本音響学会誌、47巻, 12号, 951-953 (1991)

[2] 板橋「単語音声データベース」 日本音響学会誌、41巻,
10号, 723-726 (1985)

(4)研究用日本語音声データベース

(国際電気通信基礎技術研究所：ＡＴＲ)

・セットＡ（単語音声データベース）

単語音声(8500語)、男性１０名、女性１０名

オープンリールまたはデータカートリッジ

・セットＢ（文音声データベース）

連続音声(文音声503文10000語)、１０話者

・セットＣ（多数話者音声データベース）

単語音声＋連続音声、２０話者

索　引

索　引

A
AlFresco 122
ABI 223
ACRES 128
AD−DA変換器 221
ADPCM 149
API 86, 223
ARMADA 247
ATIS 239
ATNG 101

B
Baysesの決定法 138

C
CAIシステム 50, 184
CKY法 155
CMU 22, 174, 239, 240
Cocosda Meeting 254
COC法 159
CORB 86
cosh尺度 148
CRIN 249
CV 158
CVC 159

D
DARPA 22, 174, 239
DEBUGGY 59
DECIPHER 174
Differential Modeling 58
Divergence 137
Dragon 174
DSP 221
DTW 141

E
ELIZA 119
EMアルゴリズム 146

F
FEP 221
F比 137

G
GMD 247
GUI 89, 216
GUS 120

H
HCI 216
HEARSAY−II 112
HMM 134, 181
HMM合成 194
HMM分解 194
HMM法 222

I
ICSLP 22, 239
IDAS 122
ill−formed 187
IMA 86, 233
Imposter 134
Inside−Outsideアルゴリズム 106
ITS 58

K
KL展開 140
Knock−out法 138

L
LDA 248
LDC 240
LIMSI 249
Lincoln Lab. 174
long distance dependency 204
LPCケプストラム係数 140
LR法 104
LSP方式 161

M
MOP 57, 109

N
N−bestアルゴリズム 196
NENE 127
NTT 173

O
OMA 86
OMG 86
ORB 86
Overlay Modeling 58

P
Perturbation Modeling 58
PHP 225
PLP 149

Q
Quick Time 85

R
RASTA処理 149
RSRE 247

S

SHRDLU 121
SLS 22, 239
Speaker Identification 134
Speaker Verification 134
Speaker Authentication 134
SPHINX−Ⅱ 174, 240
SRI 22, 174, 239
SUNDIAL 245

T

TRIE構造 99
TSS 86

U

UC 120

V

VCV 159
Voice Across America 240
VQ歪み 145

W

Wall Street Journal 174
Windows MME 85
Wizardシステム 199
Wizard方式 199
WUSOR−Ⅱ 58

あ

アクセント核 155
アクセント成分 154
アニメーション 214
アプリケーションシステムの分類 27
アプリケーション開発者 27
アプリケーションビルダー 51
アプリケーション 217

い

言い直し 187
言い淀み 187
一般化予測LRパーザ 195, 201
移動体通信 213
意味カテゴリ 185
医療診断書 165
インタフェースの関係 38
インフラ 227

え

エージェント 222
エディタ 217
エマージェント (emergent) モダリティ 127
エルゴードHMM 144
遠隔マイク 180

お

オープン環境 87
応用開発支援 27
オブジェクト化 233
オブジェクト変換 233
オペレーティングシステム 223
重み付きユークリット距離 140
音韻系列のマージ 204
音韻バランス文 146
音響セグメント 142
音響伝送歪み 180
音響モデル 227
音声インタフェース開発支援ツール 28, 89
音声応答機能 21, 97
音声規則合成 15, 21, 91, 153
音声合成 21
音声合成システム 70
音声合成 97
音声合成ドライバ 221
音声制御クレーン 44
音声ダイアル 226
音声タイプライタ 134
音声対話 11, 49
音声対話アプリケーション開発支援ツール 28
音声対話機能 33
音声対話システム 123
音声DB検索システム 231
音声テンプレート 219
音声入力券売機 44
音声認識・理解システム 11, 49
音声認識の形態 64
音声認識・理解応用システムの分類型 89
音声認識FEP 221
音声認識／合成サーバ 221
音声の継時変化 141
音声モデル 219
音声リモコン 177
音声ワープロ 16, 46, 92, 165
オンライン適応機能 179
オンライン適応手法 181

か

回帰係数 140
外国語会話 50
解釈・計画型 11, 42, 49
解釈型 28, 42
回線雑音 148
概念からの音声合成 21, 97
開発支援ツール 49
学習者モデル 188
カクテルパーティー効果 234
格フレーム 107
確率文脈自由文法 105
家事ロボット 177
仮想現実感 125

可変語彙 9，37

き

キーボード 216
キーワードの自動抽出 233
キーワード発声 8，35
擬音韻 142
機械主導型 11，39，49
教育型 11，28，42，49
教師無し話者適応 181
業務の自動案内 192
銀行の振込や残高の照会 192

く

組み込みインタフェース条件 29
組み込みインタフェースの標準化 30
クライアント 221
クラスタリング 144
群遅延 149

け

ゲーム型 184
警告機能 29，71
計算機用音声インタフェース 17，93
携帯情報機器 166
形態素解析 153
携帯端末 17，93，213
携帯電話 225
ケプストラム係数 140
ケプストラム方式 162
言語学習システム 184
言語指向 117
言語情報 136
言語処理 195
言語表現の訓練 185
減衰歪み 148

こ

コードブック 145
コードレス 225
語彙 8，36
航空管制システム 250
航空機のフライト情報案内 192
口述筆記 170
高信頼化 213
高信頼性化 34
高速入力機能 33
候補列のマージ 196
項目入力型 11，28，41，49
高齢者 90
黒板モデル 111
個人カード 59
語数 9，37
孤立単語音声認識 177
孤立単語発声 8，35
混合ガウス分布 146
コンフュージョン・マトリックス 72

さ

最悪状態 73
最小予測誤差 147
最長一致法 154
ささやき声 229
詐称 134
詐称者受理 134
雑音処理 194
雑音処理技術 20，96
雑発声 8，36
サブワード 142

し

視覚 125
歯科検査記録システム 44
時期的変動 140
思考の中断 185
時刻表や到着の案内サービス 192
自信の無い言い方 188
システムの評価 197
システム評価法 69
視線 131
自然言語インタフェース 119
自然発話 182
自然発声 8，36
自然発生的音声 220
自動車電話 225
指導方略 188
自動ラベル付 134
シナリオ 197
島駆動法 112
シミュレーション会話 50
シミュレーション型 184
自由発話 8，35
自由発話音声 18，94
自由文書作成 166
受聴タイミング点 159
出力ユニット 147
主導権 39
使用語彙選択支援ツール 71，90
障害者 90
情緒性情報 136，179
情報案内システム 16，93
情報案内 191
情報家電 223
情報量 38
省力化 34
触覚 126
シングルモーダル・インタフェース 117
人工知能要素技術 15，91

す

スクリプト 52，56，108
スクリプトカード 53
ステレオタイプモデル 110
スペクトル等化処理 141

せ

正準判別分析 143
声帯 137
声道 137
声道アナログ方式 161
セキュリティ 213
説明型 184
接話マイク 180
線形回帰係数 146
線形直交変換ベクトル 142
線形予測分析 139
選択機能 29，71
選択制限 107
選択的線形予測分析 148
宣伝効果 35

そ

騒音環境 9，37
相互主導型 39
操作方法の標準化 30
相談型 42
双方向主導 186
即時化 33
速記メモ 166
ソナグラフ 134

た

大語彙音声認識 17，94，194，220
耐雑音性 181
対話 229
対話音声 180
対話音声処理 18，94
対話モデル 196
ダウンサイジング 223
ダウンロード 219
多項目入力 32
多次元ガウス分布 136
タスク限定型 39
タスクの限定 39
タスクの設計 67
タスクフリー型 39
タスクモデル 51，56
多段階認識 147
単語対文法 101

ち

逐次判定法 147
チケットぴあ 209
知識源 188
知的言語指向インタフェース 119
知的ＣＡＩ 16，92
チャートパーザ 195
チャートパーザ法 155
チャート法 103
調音結合 140，168
聴覚 126
長時間平均スペクトル 142
調整子 130
超大語彙音声認識 220
直接操作 117

て

データベース 222
提案機能 29，71
低価格化 34
定型文 165
丁寧発声 8，36
ディクテーション型 28，43
ディジタル符号化音声 148
適応子 130
テキスト－音声変換 21，97
テキスト依存型話者認識 137
テキスト独立型話者認識 137
テキスト独立型の話者認識 179
デジタルネットワーク統合サービス 14，83
デファクトスタンダード 85
デルタケプストラム 141，145
電子手帳 225
電子秘書 225
電動ベッド 177
点ピッチ法 160
テンプレート 144
電話音声 148
電話回線サービス 44
電話回線入力機能 32
電話口述筆記 166
電話番号案内サービス 197
電話番号案内 200
電話番号の案内 192

と

等誤り率 136
同義語 72
統計言語情報 169
統計的なＬＲ構文解析手法 246
同定型 28，41
動的計画法 141，154
動的特徴 141
登録話者 134
特徴ベクトル 136
特定業務文書作成 165
特定語彙 9，37
特定話者 8，36
トップダウン解析法 155
ドリル型 184

な

ナビゲーションシステム 17，93

に

肉等等級入力システム 44
２次臨界制動系 141
日本語ｄｉｃｔａｔｉｏｎシステム 173

ニューラルネットワーク　147，222
任意語彙　9，37
人間主導型　39
認識誤り　71
認識対象語彙セット　29

ね
ネットワーク　221
ネットワーク構造　99
ネットワークモデル　112

の
ノイズキャンセル　234
ノウハウ的留意事項　11，29，63
ノンバーバルインタフェース　129

は
バージイン　194
パーソナルコンピュータ　216
ハーバード大学　22，239
配送仕分システム　43
ハイパーテキスト　214
ハイパーテキスト構造　217
ハイパーメディア構造　217
ハイパーモーダルインタフェース　127
拍同期点　159
パスワード　177
発声変形　229
パソコン通信センター　218
発話環境処理　20，96
バックスペースキー　220
バックワードのスコア　196
発声訓練　185
発声形式　8，36
発声単位　7，35
発声方法　194
話し言葉　229
破裂音　144
バンキング　134
反射係数　148
ハンズフリーマイク　172，218
ハンドフリー音声認識電話　45
判別分析　142

ひ
ビームサーチ　228
ビーム探索法　112
鼻音　144
歪み・重なり尺度　146
非接触制御機能　32
非直線歪み　148
ピッチ　137，182
ピッチパターン　154
ビデオ予約リモコン　44
（人←→機械←→人）型　38
（人←→機械）型　38
ビューア　217
評価　29
評価手法　90
標識　130
標準化　29，90
表情　131
表情合成システム　124
表情つき音声対話システム　122
標本化周波数　140

ふ
フィリップスアーヘン研究所　248
フェイルセイフ　178
福祉ロボット　16，92，177
複数特定話者　8，36
複数のマイクロフォン　194
付属語アクセント結合　156
付帯語　187
不特定話者　8，36
不特定話者方式　71
不特定話者音声認識　19，95，193，222
部分空間　144
部分バージング　187
プラン・ゴールモデル　110
ブリストル大学　246
プルダウンメニュー　218
フルテキストサーチ　214
プロソディー　140
文ー音声変換　86
分散比　137
文節数最小法　154
文節発声　8，35
文テンプレート　101
文脈自由文法　101，155

へ
平均基本周波数　139
平均ベクトル　136
ベクトル量子化　134，146，182
ヘッドセット　219
変換・翻訳型　43
ペン入力コンピュータ　216
ペン入力　225

ほ
ホルマント合成方式　161
ホルマント周波数　137
ホログラフ　214
本人否却率　134

ま
マーカバッシング　113
マイクロホン　9，37
マウス　216，218
マクロ命令　177
摩擦音　144
マハラスビス距離　140
マルチウィンドウシステム　217

マルチテンプレート 181
マルチテンプレート法 222
マルチメディア 22, 83, 98, 191, 231
マルチメディアデータベース 17, 231
マルチメディアデータベース検索システム 93
マルチメディアPC 85
マルチモダール 22, 98, 180, 191, 222
マルチモーダル・インタフェース 121, 172
マルチモーダル化 212
マルチモーダル対話 234
マンマシンインタフェース 178
マンマシン対話 177

み
ミクロな命令 177
未知語処理 185
未知語入力 169
道順等の地理案内 192
身振り 130

む
無意味語 187
無声部 146
無線ネットワーク 225

め
明瞭度 220
命令型 11, 28, 41, 49, 182
メルケプストラム 149
メルケプストラム方式 162

も
盲導犬ロボット 178
目標達成 75
文字認識技術 216
モダリティ 117

ゆ
ユーザインタフェース 15, 21, 68, 75, 91, 98
ユーザの満足度 75
ユーザフレンドリー化 34
ユーザモデル 51, 57, 109
有限状態オートマトン 53
有声部 146
ユニフィケーション技術 195

よ
用語セット 72
予測ニューラルネットワーク 147
予備選択 169
予備選択法 228
読み上げ発声 8, 35
予約システム 17, 93

り
力覚 126
リジェクト機能 134
リフター 178
旅行相談 192
臨界制動モデル 160

る
ルールベースシステム 53
類義語辞書 73
類似度マトリックス 73

れ
例示子 130
レベル変動 180
連続音声認識 18, 94
連続音声認識アルゴリズム 196
連続数字認識 220
連続単語発声 8, 35
連濁 158

ろ
ロジカケンブリッジ 245
ロンバート効果 21, 229

わ
ワークステーション 216
ワードスポッティング 181, 233
ワープロ 218
話者 8, 36
話者識別 96, 134
話者照合 134
話者情報 136
話者適応 8, 19, 36, 95
話者認識 20, 96, 228
話者認識技術 15, 91
話調成分 154

- 禁無断転載 -

04-R 012
音声の知的処理に関する調査研究報告書

発　行　平成５年３月
発行所　財団法人　日本情報処理開発協会
東京都港区芝公園３丁目５番８号
機械振興会館内
電話（０３）３４３２－９３９０