マルチモーダル インタフェースの要素技術
Component Technologies for Multimodal Interface

金澤 博史 H. Kanazawa　松浦 博 H. Matsuura　辻本 修一 S. Tsujimoto

マルチモーダル インタフェースを構築するためには，音声，文字，画像などのメディアの単なる入出力ではなく，これらのメディアを介してやりとりされる情報の理解，生成，提示技術が重要である。さらに，実際の応用場面ではさまざまな利用状況を考慮した頑健な処理方式の開発が必須(す)となる。例えば，音声処理では耐雑音認識や高音質音声合成，文字認識では罫(けい)線除去や崩し文字への対処，画像では動画像への対処，提示技術ではシステムからユーザへの自然な情報の伝達などである。これらの課題に対するアプローチが実用化へのキーとなる。

When constructing a multimodal interface system, it is important not only to develop simple media input and output technologies, but also understanding and synthesis technologies for the contents communicated through speech, characters, images, and so on. Further, consideration should be given to the development of robust processing techniques for real-world applications. For example, noise robustness, elimination of constraints on the users, and naturalness are problems that must be resolved. Achieving robustness is a key approach in realizing a practical multimodal interface.

1 まえがき

人間は，コミュニケーションにおいて言語，音声，ジェスチャ，表情などさまざまなメディアを駆使して，意志の伝達が円滑に行われるように相手に働きかける。そこでは，単なるメディアの入出力ではなく，メディアにより表現される内容の理解と生成が重要となる。このことは，計算機とのインタラクションにおいても同様である。すなわち，計算機が人間と円滑なコミュニケーションを行うためには，さまざまなメディアを介して人間から発せられる情報をいかに取り出し，理解し，応答するかがキーとなる。この意味で，マルチモーダル インタフェース技術は，図1に示すように計算機に人間と同じような五感を与え，それらのセンサから得られる情報を正しく理解し，適切な応答を行い，円滑な対話を行うための知識を与えることに相当する。

当社では，マルチモーダル インタフェースに向けたさまざまな要素技術の研究開発を行っているが，ここでは特にメディアの認識・理解および生成技術を中心に述べ，それらを利用したマルチモーダル インタフェースの事例を紹介する。

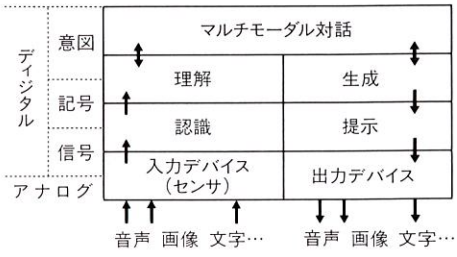

図1．マルチモーダル インタフェースを構成する要素技術　それぞれの処理間で相互に情報をやりとりしながら，ユーザの入力を理解し，適切な応答を生成する。

Component technologies for multimodal interface

2 音声処理技術

2.1 音声認識

最近，音声言語による計算機への入力が徐々に実用化されてきており，音声の応用場面が広がっている。当社ではさらなる応用拡大のために，環境雑音や発話様式の違いに対処し，不特定ユーザが自由に音声を利用できるような認識方式の開発を進めている。

耐雑音性の向上のために，入力音声を認識用の音声特徴ベクトルに変換する特徴抽出処理と，認識処理の2段階で雑音に対処する。特徴抽出部では，雑音による時間・周波数成分のさまざまな変動を表現したセグメント単位のマッチングにより，雑音に影響されない安定した特徴ベクトルを求める。さらに，認識部で用いる辞書を雑音重畳音声により学習し，特徴抽出部で吸収しきれない変形に対処する。これにより，高い耐雑音性を実現することが可能となった。

また，発話様式の制限を緩和するため，キーワード スポッティングによる話し言葉を対象とした意図理解方式を開発した。ここでは，話し言葉からあらかじめキーワードと

して定められた単語だけを抽出し，キーワードのつながりからユーザの発話の内容を理解する。これにより，不要語だけでなく，語順の逆転や言いよどみなどの話し言葉特有の現象にも対処でき，ユーザの負担を大幅に軽減することができた(1)。

2.2 音声合成

パソコンやインターネットの普及を背景に文章を音声に変換する音声合成技術に対する需要が増大している。しかし，一般ユーザが音声合成を違和感なく利用するには，音質や自然性など解決すべき課題が少なくない。特に，音質の問題は自然な声で発話内容を明確に提示するうえで重要であり，その解決のためLPC (Linear Predictive Coding) 分析残差駆動方式による合成技術を開発している(2)。

図2は音声合成方式の全体構成である。この方式は，あらかじめ収集した自然音声のデータベースから合成に用いる代表音声素片を自動的に学習する機構をもち，学習で得られたCV，VC (C：Consonant，V：Vowel) 素片のピッチや継続時間長を制御した後，これらを接続して音声信号を生成する。代表素片の学習は，合成音と自然音声とのひずみの総和を最小化する規範で，データベースから素片を自動的に選択することで実行される。

この方法により，従来，音質劣化の主要因であったピッチ変更に伴うひずみが激減し，明りょうで肉声感豊かな音質を実現できるようになった。また，この方式では音声素片のパラメータの操作により声質の変更が容易であるとともに合成素片辞書のサイズが小さくて済み，携帯情報端末やカーナビゲータなどへの組込みも容易である。

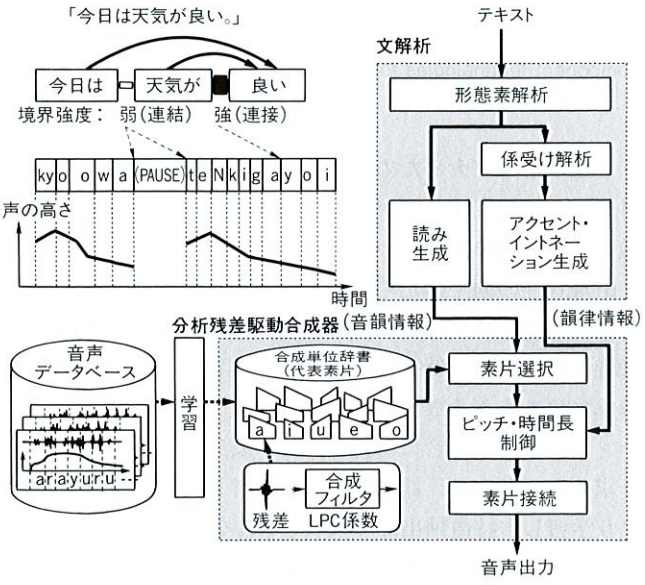

図2．音声合成方式の全体構成　最適な素片選択のための学習と係受け解析の精緻(ち)化により高音質を実現している。

Overview of high-quality text-to-speech synthesis

3 文字認識技術

3.1 フォームリーダ

フォームリーダは，銀行振込券のような帳票の枠内に記入された文字を読み取るためのものである。従来，専用の帳票だけを対象としていたが，最近では，既存の帳票も扱えるようになってきた。しかし，①罫線の位置ずれや伸縮などの影響を受けやすい，②ユーザが文字を記入する際に罫線に接触しないように注意する必要がある，などの制約があった。

そこで，①に対しては入力帳票から抽出した罫線情報に対し，あらかじめ登録されている帳票の特徴を記述した幾何モデルを適用し，グラフ理論に基づく新たなマッチング法を導入し，入力帳票の大局的/局所的変動を吸収することができるようになった(3)。また，②に対しては，文字と罫線との接触部分付近の画像の輪郭を追跡し，文字と罫線候補との接点を切断点として，切断点どうしを組み合わせて最適な切断線を決定する手法を開発した。これにより，文字の輪郭を正確に残して罫線だけを除去することができ，高精度な文字認識が可能となった。

3.2 オンライン文字認識

ペンを用いたオンライン文字認識は携帯端末などに広く利用されており，当社ではディスプレイに装着したタッチパネルに指で直接触れて文字を書く指書き文字認識を開発した。図3は，ATM (Automatic Teller Machine) の受取人名の入力に指書き文字認識を適用した例である。指書きならば，すでに入力装置として用いられているタッチパネルとのシームレスな操作が実現できる。

当社で開発したオンライン文字認識方式の特長は，崩し字や続け字の柔軟な認識と，ていねいに書かれた文字の高精度での認識を両立させた点にある。一般に，ていねいに

図3．ATMにおける指書き文字認識の利用　指で書いた筆跡が黒く表示される。訂正，確認などのタッチ入力との親和性のよい操作環境を実現している。

Finger-written character recognition on automatic teller machine

書かれた文字を正確に認識するためには，崩し字や続け字を高頻度でリジェクトする必要がある。逆に，崩し字や続け字を許容すると，ていねいに書かれた文字を誤認識する危険性が高くなる。当社では，崩し字や続け字に強い方式とていねいに書かれた文字に強い方式を併用した統合アルゴリズムを開発し，ラフな入力で認識誤りがあったときに，ていねいに書き直せば確実に入力できるようになり，何度書き直しても入力できないという従来の問題を回避できた。

4 画像処理技術

画像処理技術のなかで，インタフェースへの応用という観点からもっとも重要なものの一つが顔認識である。顔画像を用いてユーザの識別が行えれば，個人情報を利用したさまざまな応用を考えることができる。

顔認識の処理の流れを以下に述べる。まず，顔領域の抽出では，顔の濃淡パターンから作成した辞書を用いて全画面探索を行い，類似度がしきい値以上となる箇所を顔領域として検出する。次に，分離度フィルタ[4]を用いて特徴点候補を検出し，4点（両目，鼻腔）からなる目鼻特徴点を抽出する。さらに，目鼻特徴点を基準に顔領域の正規化を行い，あらかじめ個人ごとに作成した顔画像の濃淡パターン辞書との類似度を求め，最大類似度をもつ人物を識別結果とする。

この手法では，動画像を用い，複数の連続した画像に対する識別結果を統合することで，性能向上を図っている。

5 情報提示技術

5.1 擬人化エージェント

計算機との自然な対話を行うための情報提示技術に，擬人化エージェントがある。これは，計算機を仮想的な人間とみなして，人間を擬したキャラクタを表示させることにより，ユーザの心理的負担を軽減するとともに，計算機の内部状態や提示すべき情報をユーザに円滑に伝達する技術である。例えば，マルチモーダル秘書エージェントシステム[5]では，口や目，まゆの動き，顔の向きや身振りなどを細かく制御した三次元CG（Computer Graphics）を用いている。

図4に示すように，対話がうまく進行しているときには笑顔，対話の停滞時には悲しい表情をするなど，ユーザに視覚的に現在の対話の状態を示し，円滑な対話の進行を助けている。また，規則合成の音声出力と同期して口を動かすとともに，発声内容をテキスト表示することにより，音声メディアの一過性の欠点を補っている。この擬人化エージェントは，音声とCGを統合化することで，計算機との対話を人間どうしの対話にいっそう近づけ，ユーザの拒否感を軽減し，システム全体のパフォーマンス向上に大きく貢献した。

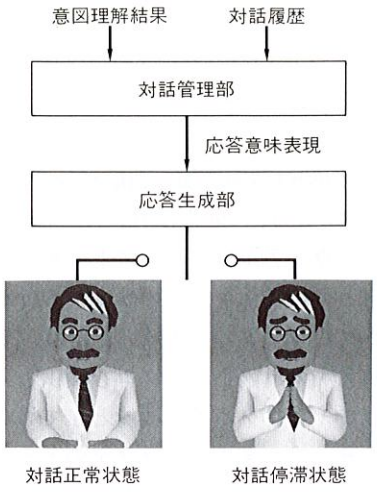

図4．三次元CGによる擬人化エージェント　意図理解結果と対話履歴から得られた応答意味表現に基づき，システムの状態（対話正常，対話停滞）をCGの表情で伝える。
Animated human-like agent using 3-D computer graphics

5.2 簡易アニメーション表示

ユーザにシステムの操作手順を説明したり，動画情報を提示するために，簡易アニメーション表示技術がある。この技術はアニメーションのキーフレームとなる画像を線画で表現し，複数次元ベクトルの補間技術を線画の補間に応用することによって，キーフレームからさまざまな画像をリアルタイムに生成・表示するものである。画像を補間生成することから，小数の画像からアニメーションの生成ができるという特長がある。また，リアルタイムに画像を生成できることから，インタラクティブ性に優れ，表示画像の設計工数を大幅に短縮でき，プロトタイピングの高速化を図ることができる。

6 マルチモーダル インタフェースの事例

6.1 HIウェア

HIウェアは，文字認識や音声認識などのHI（ヒューマンインタフェース）機能を利用者が簡単に利用できる環境を提供する。このソフトウェアはWindows95®(注1)上で動作し，さまざまなシステム構成に柔軟に対応できるよう，クライアント／サーバモデルを採用し，HIウェア利用のアプリケーションを簡単に構築できるようにした。

例えば，HIウェアを利用した英文のファックス文書の読上げシステムでは，文字認識，機械翻訳，音声合成を組み合わせて，ファックスで配送された文書を計算機に取り込

（注1）　Windows95は，Microsoft社の商標。

み，日本語のテキストに変換し，音声で出力することができる。HIウェアを用いることにより，ユーザが複数のアプリケーションを個別に立ち上げることなく，さまざまに組み合わせて高度な処理を行うことができる。

6.2 パソコン向け音声認識・合成インタフェース

当社のパソコンには，“東芝音声システム”として音声認識・合成ソフトウェアがプリインストールされている。音声合成の応用としてはテキストの読上げのほか，キーボードから入力されたコードの読上げ，表計算，ワープロソフトウェア中のテキスト読上げマクロを提供している。

また，擬似キャラクタボイス機能として男女声の基本周波数を変化させ，アニメのキャラクタのような声を合成する機能も付加している。認識ソフトウェアは，デスクトップおよびノートタイプの内蔵または外付けマイクで利用でき，文字列登録によって認識語彙(い)の追加変更が可能である。Windows95®コマンドやユーザ設定コマンドを実行でき，さらに，認識対象単語の前後に，「えーと」などの不要語を付加しても認識できるようワード スポッティング機能を加え，発声様式に対するロバスト性を高めている。

6.3 マルチモーダル秘書エージェント

マルチモーダル秘書エージェントは，人に代わって問合せに答える知的な情報検索システムである（図5）。このシステムでは，各種のメディア理解技術を統合し，音声やジェスチャ，キーボードによるマルチモーダル入力を可能としている。複数メディアによる入力からユーザの意図を抽出するためには，各メディアの処理結果をいかに統合するかが問題となるが，その方策として新たな入力統合方式を開発し実装している。

統合の際には，各メディアの複数の認識候補から，仮説推論により，過去に行われた入力統合結果を参照し，矛盾のない候補の組合せを絞り込み，解釈処理を行う。これにより，複数候補の組合せによる候補の爆発を抑え，高速にかつ精度よく正しい解釈を導くことが可能となる。

また，出力は擬人化したエージェントのCG画像，合成音声，応答文テキスト，グラフィックスによるマルチモーダル出力であり，各場面の応答結果画像に対してはジェスチャによる指定を行えるようにした。入力解釈に失敗した場合は，再入力の回数に応じて表情を徐々に変化させたり，合成音の韻律を変化させるなどして非明示的にシステムの状態を伝達できるようにした。

図5．マルチモーダル秘書エージェント　音声入力とタッチ入力を同時に行い，それらを統合した解釈をすることができる。
Multimodal interface agent

7 あとがき

メディアの理解，生成，提示技術を中心に当社のマルチモーダル インタフェース要素技術の一部を紹介した。マルチモーダル インタフェースが，ポスト GUI (Graphical User Interface) として真にユーザに受け入れられるためには，メディアの理解・生成の枠組みをいかに構築するかが重要となる。そして，ユーザがなんら意識することなく，自然に本来の目的を達成できるインタフェースこそがわれわれの目ざす姿である。

文　献

(1) 竹林洋一：音声自由対話システム TOSBURG II――ユーザ中心のマルチモーダル インタフェースの実現に向けて――，電子情報通信学会論文誌 D-II，J77-D-II，8，pp.1417-1428 (1994)
(2) T. Kagoshima, et al: Automatic Generation of Speech Synthesis Units Based on Closed Loop Training, Proc. ICASSP97, pp.963-966 (1997)
(3) 石谷康人：モデルマッチングによる表形式文書の理解，PRU-94-34 (1994)
(4) 山口　修，他：分離度特徴を用いた顔画像解析――目，瞳の検出――，情報処理学会第52回全国大会，2-187 (1996)
(5) 中山康子，他：マルチモーダル秘書エージェント，東芝レビュー，52，5，pp.25-28 (1997)

金澤　博史　Hiroshi Kanazawa
研究開発センター 関西研究所研究主務。
音声認識・音声対話技術の研究開発に従事。電子情報通信学会，日本音響学会会員。
Kansai Research Lab.

松浦　　博　Hiroshi Matsuura, D.Eng.
マルチメディア技術研究所 開発第六部グループ長，工博。
音声処理・ヒューマンインタフェース技術の研究開発に従事。電子情報通信学会，日本音響学会会員。
Multimedia Engineering Lab.

辻本　修一　Shuuichi Tsujimoto
研究開発センター 情報・通信システム研究所研究主務。
マルチモーダル インタフェースの研究開発に従事。電子情報通信学会会員。
Communication & Information Systems Research Labs.