2M－8

音声入出力とタッチパネルを用いたマルチモーダル対話システムの試作

松浦　博，正井康之，原　義幸，新田恒雄
（株）東芝　情報処理・機器技術研究所

１．まえがき

音声認識を一般のユーザーが利用する社会システムに応用するためには，不特定話者に対する認識性能を高めることや[1]自由発話を認識すること[2],[3]だけでは十分ではない．ユーザーがシステムに向かった時に，戸惑うことなく気軽に発声・操作することのできるヒューマンインタフェース環境を提供する必要がある．

最近，特にユーザーに満足感を与えるように配慮したシステム，さらに進めてユーザーオリエンティドなシステムの製品化が社会的にも要請されている．本報告で述べるマルチモーダル対話システムでは，入力手段および出力手段の双方をマルチチャンネル化すること，ユーザーの動作をセンサにより把握すること，対話的に操作を行うことによって，この要請に答える．

２．システム構成

今回，試作したシステムは，一般のユーザーを対象に，東京駅周辺の情報案内を行う．システムの概略構成を図１に示す．入力手段に音声認識ユニットとタッチパネルによる直指（直接指示），出力手段に規則音声合成ユニットとディスプレイを採用した．主な機能について以下に説明する．

２．１　音声認識ユニット

音声認識は不特定話者の自由発話を対象とするＳＭＱ／ＨＭＭに基づく方式[2]を採用した．この方式はこれまで，単独発声された１０００単語の大語彙認識に適用され，認識率９６．３％を得ている[1]．音声認識ユニットの特長を次に示す．

①ＳＭＱ／ＨＭＭは話者適応の必要のない純粋の不特定話者音声認識であり，老若男女が入れ代わり使用する社会システムには最適である．

②自由発話を対象とするため，不慣れな人による，「えーデパートはどこかな」あるいは「トイレトイレ」などといった発話からキーワードのデパートやトイレをスポッティングでき，高精度に認識する．

③不特定多数のユーザーを想定し，多様な表現をカバーしている．（例：駅／東京駅／ＪＲの駅）

④32個のキーワードおよび５個の補助単語と17個の未知語を含む，計54個の単語セットから95種類の複合語を作成して評価実験を行ったところ，キーワード認識率で91.1％を得た．

⑤音声入力はHiFiマイクを組み込んだハンドセットを使用しているため，ユーザーに馴染みやすい．

２．２　タッチパネル

タッチパネルは，一般のユーザーにも使い方の習得が容易で，しかも表示と選択項目が一致しているため使い勝手が良い．しかし，多項目の選択には不向きである．一画面の項目数は８個程度に押さえるのが適当であるとされ[4]，それ以上は一般に次画面表示と項目を階層化することが推奨されている．しかし，次画面表示や階層化は操作時間の増大につながる．また階層化においては階層がどの様に構築されているかを誰にでも自然で分かりやすく，しかも共通の認識が得られるようにするのは容易ではない．このため，多項目の選択には「直接アクセス可能」な音声認識が適している．タッチパネルを共用することの長所を以下に示す．

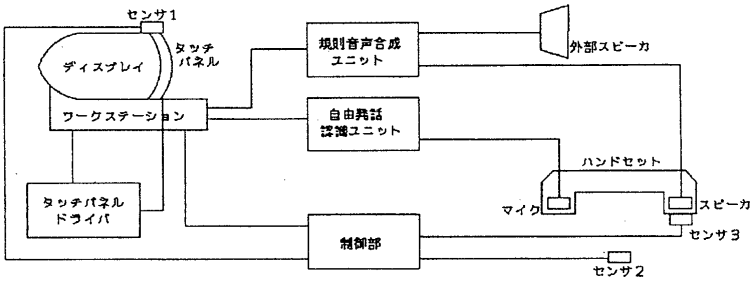

図１　システムの構成

Multimodal Dialogue System Using Speech Input/output Device and Touch-screen.
Hiroshi MATSUURA, Yasuyuki MASAI, Yoshiyuki HARA and Tsuneo NITTA
Information Systems Engineering Laboratory, TOSHIBA Corporation

①既存の案内システムと同じ感覚で利用できる．
②音声認識の不確実性を補うことができる．
③音声認識と協働で多彩な機能を果たす．
例：「デパート」と発声し，複数のデパートを表示させ，タッチ入力で選択するなど．

２．３　規則合成ユニット

規則合成はディスプレイ表示だけでは不十分な次のような場面，すなわちユーザーへの使い方の指示，使い方が適切でない際の警告，および情報の提供を行う．規則合成利用の特長を以下に示す．
①従来，音声メッセージは特定のアナウンサの発声を録音し，この音声データから必要な箇所を注意深く切り出して登録する必要があった．規則合成では仮名漢字コードで入力した文章から音声への変換を自動的に行うことができ，案内文の事前作成が容易である[5],[6]．
②時事刻々変化する情報を即座に音声化して提供するのに適している．
③音声入力時には，ハンドセット内のスピーカから音声を出力するので，認識への影響が少ない．（利用者がハンドセットを手にとるまでは，外部スピーカから出力する．）

２．４　センサ

ユーザーの情況を知ることは対話システムを構築する上で，重要である．以下に図１の複数の接近を検知する（光電）センサの役割を示す．
①センサ１はユーザーがシステムの前に近付いたことを検知する．
②センサ２はユーザーがハンドセットを手に取ったことを検知する．
③センサ３はユーザーがハンドセットを耳に当てたことを検知する．

３．システムの動作

システムの動作の概要を述べる．
①利用者がシステムの前に来ると，センサ１がこれを検知し，次にシステムは，図２のように利用者に対してハンドセットの持ち方を示す．同時に「受話器を耳に当てて下さい．」という指示を規則合成音で外部スピーカから出力し，不馴れなユーザーに音声入力の操作方法を指示する．
②次に，ハンドセットを手に取るとセンサ２が検知して，図３の画面を表示する．
③続いて，ハンドセットを耳に当てたことをセンサ３が検知すると，ハンドセット内蔵スピーカから「希望の場所を発声して下さい」という指示を音声出力する．
④案内対象の単語，たとえばデパートを発声すると，

図２　ハンドセット保持の案内画面

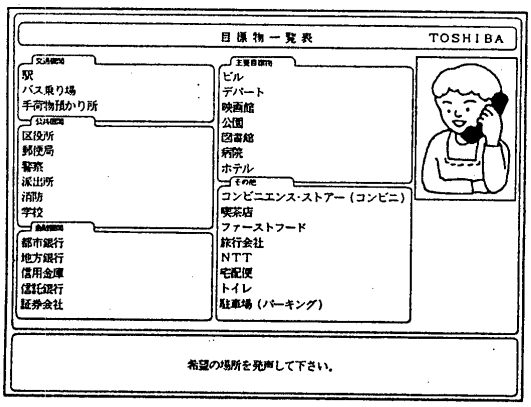

図３　発声案内画面

東京駅周辺のデパートが地図と共に複数表示される．
⑤個々のデパート名の表示を指で触ると，「本日は定休日です．」などの情報を規則合成音で得ることができる．

以上述べたようにユーザーはシステム主導であるが，規則合成と表示に従って，ハンドセットを保持し，続いて発声とタッチ入力を局面に応じて適切に行うことによって，必要な情報を素早く手軽に得ることができる．

４．まとめ

ユーザーの使いやすさを念頭に，情報案内システムをマルチモーダル対話システムとして試作した．今後は自由発話音声認識の大語彙対応，マルチモーダル機能の充実に取り組む予定である．

文献(1) 松浦ほか，音学講論，1-P-25(1992-03) (2)Y. Masai, etal, ICSLP-92(1992-10) (3)貞本ほか，人知全大，17-5(1992-6)(4)Ben Shneiderman，ユーザーインタフェースの設計 (5) 原 ほか，信学全大，A-6(1992-3)(6)小林ほか，情処全大7N-6(1992-3)