音声言語情報処理 12－13
（1996. 7. 27）

携帯端末に対応した音声対話インタフェースの検討

城塚 音也，山田 達司†，小泉 宣夫
NTTデータ通信株式会社 情報科学研究所，マルチメディア技術センタ†
e-mail:sirotuka@lit.rd.nttdata.jp

あらまし

我々はデスクトップ端末と比較してグラフィカルユーザインタフェースの有効性が限定される携帯機器への音声対話インタフェースの導入を検討している。

オンラインサービスを対象に携帯機器での音声対話の実現方法を考えると、音声認識および音声合成を行なうサーバーを外部に用意し、携帯機器が音声/データ通信によりサーバーを利用するC/S構成が有効である。また、ユーザの持つユーザインタフェースが多種多様なモダリティを持っていると仮定すると、ユーザが享受するサービスの質（QOS）の総和を高めるにはユーザの所持する機器に合わせてサービス対話に使用するモダリティ、対話モデルを変えることが有効である。我々はC/S構成を基本に、複数の種類の携帯機器により使用が可能であり、携帯機器のもつモダリティに応じた対話戦略を行なう対話型サービスサーバ Dialogue Engine を提案する。

和文キーワード 携帯端末，音声対話インタフェース，対話モデル，QOS制御

A Study on Spoken Dialogue User Interface for Mobile Devices

Otoya Shirotsuka, Tatsusi Yamada†, Nobuo Koizumi
NTT DATA COMMUNICATIONS SYSTEMS CORPORATION
Laborator for Information Technology, Multi Media Technology Center†

Abstract

We investigate the feasibility of applicability of spoken dialogue user interface (SDUI) to mobile devices which graphical user interface is not effective compared to desktop machines. For online services, realization of SDUI by server client architecture is reasonable, in which mobile devices communicate speech processing via speech and data lines. To maximize the sum of the Quality Of Service (QOS) which users receive, a selection mechanism of communication modalities and dialogue model according to their manipulating devices is desirable. Therefore we propose an dialogue-based service server Dialogue Engine to realize such flexibile funcitons.

英文 key words mobile device, spoken dialogue interface, dialogue model, QOS control

1 はじめに

電子手帳やPDA (Personal Digital Assistant) と呼ばれる携帯性を重視した携帯端末が普及している。液晶の画面をもち、ペンやキーボードといった入力手段をもつ手帳大の端末は、通信機能の内蔵化により最終的には携帯電話と融合した形態 (smart phone 等) へ進化を遂げると予想される。また、デスクトップPCにおけるマルチメディア情報処理能力は、ノート型PCにおいて実現されており、将来的にPDAレベルの携帯端末においても実現されるであろう。

音声とデータ通信機能や、マルチメディア情報処理能力を得た携帯端末は、孤立したスケジュール管理や個人情報のデータベースのみならず、マルチメディア型のＡＰや、オンラインサービス等に利用可能となる。このような将来の携帯端末のユーザインタフェース (UI) を考えると、デスクトップ向けのGUIはusabilityに問題があり、携帯端末むけの限定されたリソースを有効に使用できる独自のUIが必要であるという指摘がある [1]。われわれは、通信とマルチメディア情報処理能力を身につけた将来型の携帯端末を想定し、音声入出力機能を備えたUIの検討をおこなっている。

本報告では、携帯端末における音声入出力機能の実現方法を検討し、オンラインサービス向けのUIとして、クライアント／サーバ (C/S) 構成の音声対話機能の実現について説明する。さらに、ユーザが所持する様々な携帯機器により、可能な限り快適にサービスを提供する仕組みである Dialogue Engine を提案する。

以下, **2**.において携帯端末における音声入出力の実現方法について議論し、**3**.において, 柔軟な対話環境を実現するDialogue Engineについて述べ、**4**.では, 会議室予約サービスをタスクにDialogue Engineにより実現された携帯機器のユーザインタフェースの対話実験について報告する。そして、**5**.においてまとめと今後の課題について述べる。

2 携帯端末における音声入出力の実現方法

現在の携帯端末は、バッテリーの問題等により、強力なＣＰＵを搭載しておらず、音声等のマルチメディア情報処理機能を備えていない。

音声認識、合成を携帯端末で行う場合、その方法は大きく3つの方法に分けられる。

1. 現状のハードに専用の音声処理カードやチップを搭載する。

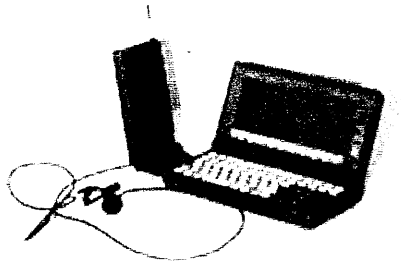

図 1: 通信機能をもったPDAのイメージ

2. 現状のハードに音声／データ通信機能を付加して認識合成処理を通信相手のサーバーに行わせる。

3. 現状のハードにAD/DAハードを搭載し、CPUを強化してソフトウェアのみで認識合成処理を行わせる。

1についてはデスクトップやノート型のWS、PC用に試作されている [2]。2はLAN環境下の計算機において検討されている [3]。3はデスクトップPC上で実現されているソフトウェア音声認識を携帯端末で行う試みであるが、現在のところ実現されていない。

2の負担の大きい処理を外部のシステムに任せるクライアントサーバ型の処理方法は通信機能を取り込みつつあり、かつ非力なCPUをもった携帯端末にとって適している。アナログ音声をサーバに送り、サーバがA/D変換、特徴抽出、パタンマッチングを行ない認識結果を返すという処理分担が考えられる。

また、携帯電話（デジタルセルラー）と融合した携帯端末の場合、音声符号化用DSPを音声通信時と切り替えて使用することにより、音声のデジタル化と特徴抽出を行うことが考えられる。アナログ音声の代わりに特徴量を送ることにより通信量を大幅に削減することができる。また、アナログ電話音声認識の問題である回線ノイズを無視することができ、認識性能の向上が期待できる。図1はPHSカードによるPDAの音声入出力機能実現のイメージである。

このようなC/S構成の音声入出力が適したアプリケーションとしてオンラインサービスが考えられる。これは、サービス自体に通信機能が必要であり、サービスセンターが音声認識、合成サーバーを管理することにより、認識合成用の辞書のメンテナンスを行なうことが簡単に行なえるためである。

そこでわれわれはオンラインサービスをタスクに携帯端末を使用した音声対話 UI の検討を行った。

3 柔軟な対話環境を実現する Dialogue Engine

マルチメディア通信において、通信路のデータ転送速度や端末の能力に応じて、再生する画像や音声の「質」を制御する、Quality Of Service (QOS) 制御が検討されている [4]。これにより、ユーザは、自分のおかれている情報環境に応じた、快適なサービスを受けることができる。

われわれは、この QOS 制御のコンセプトを拡張し、音声対話を中心にしたインタラクティブなサービスの質を情報環境に応じて制御することを考えた。人間と計算機の「対話」の質を規定する状況的要素として以下の 4 点を挙げる。

1. 端末環境（入出力 modality、cpu パワー等）
2. 通信環境（伝送速度、channel 数等）
3. サービスの種類（transaction、browsing、messaging 等）
4. ユーザの種類（初心者、熟練者等）

既存の UI の多くはこのような要素を固定的に設定しており、その結果、想定していない状況下で使用すると、急激に対話の QOS が低下してしまう。図 2 に情報環境の違いによる QOS の変化を示す。この図において、デスクトップ向けの GUI の QOS はデスクトップ環境に急峻なピークをもった分布で（A）、モバイルコンピューティングにおける携帯端末の QOS は、低いが幅広く諸環境をカバーする分布で（B）表される。

携帯環境では、通信、端末環境がデスクトップ環境に比べて流動的であり、固定的なインタフェースでは所有するリソースを生かした対話を行なうことが出来ない。対話の QOS 制御を行なうことにより、図 3 で示すような、複数のサービス AP の利用を考慮した QOS の総量をを最大化することを目指す。

このような対話の QOS の制御を可能にする仕組として Dialogue Engine を提案する。図 4 にその構成を示す。

Dialogue Engine は、情報環境に応じて、使用する端末の modality を決定する。また、UI の対話モデルの違いにより、UI の操作性が変化することから [5]、ユーザ自身やタ

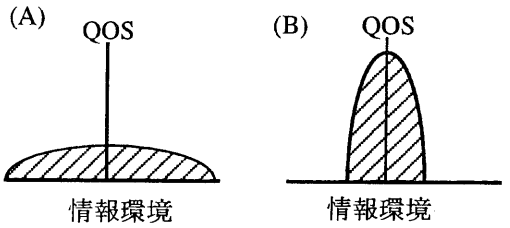

図 2: 情報環境の違いと対話の質の関係

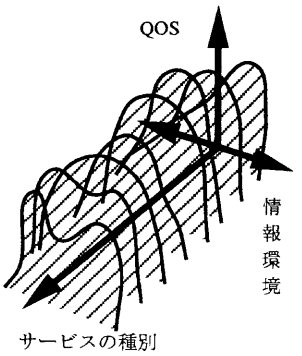

図 3: AP の違いと対話の質の関係

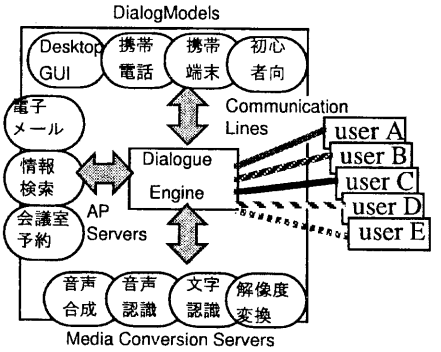

図 4: Dialogue Engine の構成

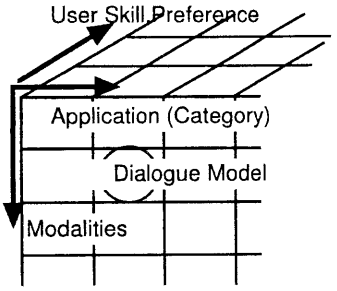

図 5: 対話モデルの決定の概念図

スクの性質、使用する modality に応じて、対話モデルの変更を行なう。図 5 に対話モデルの決定の概念図を示す。

Dialogue Engine では使用する対話モデルに基づいてユーザとの対話が行なわれ、実際のサービス処理は、各種のサービスサーバ（予約業務等）で行なわる。音声認識等の各種のメディア変換処理は、メディア処理サーバにおいて行なわれる。

また、モバイル環境での、端末環境とは、ユーザが所持している携帯機器の機能、能力であるが、ユーザが機能の異なる複数の機器を所持している場合、それらを併用することができる。これにより、使用可能な modality を増やすことができ、対話の QOS の向上に有効である [6]。

Dialogue Engine の機能として今回実現した対話の QOS を制御する仕組みは以下の 3 点である。アプリケーションとして会議室予約のオンラインサービスを用意した。

- 使用可能なモダリティの違いに対応した対話モダリティの選択
- 複数端末の併用による対話モダリティの拡張
- サービスに使用するモダリティに応じた対話モデルの切替

第 1 と第 2 は、端末、通信環境に対する柔軟な対応の仕組みである。具体的にはユーザは携帯電話と PDA のどちらか、または両方を使用することができる。その結果、各端末に備わっている対話 modality によってサービス対話が行なえる。具体的なモダリティの違いを表 1 に示す。

第 3 の仕組みは、ユーザ主導、システム主導、ユーザ主導とシステム主導の混合という対話形態の違いから分類した 3 つのモデルの選択である。ここで実現した混合モデルは、人間と同様の自由な turn-taking が行なえるモデルではなく、ユーザ主導を主として、以下のようにユーザの次行動が 1 つに予測されている場合にユーザから対話の主導権を奪い、予測に基づいてシステムがユーザに問いかけをおこなう対話モデルである。

- 対話のスタート時：サービスの利用目的を聞く
- 予約の条件がすべて揃った時：予約を行なうかどうか聞く
- 予約の条件に該当する空き部屋が 1 つしかない時：予約を行なうかどうか聞く
- キャンセルの条件に該当する予約が 1 つしかない時：キャンセルを行なうかどうか聞く

	携帯電話	PDA	併用
音声入力	○	×	○
キー入力	×	○	○
音声出力	○	×	○
画面出力	×	○	○
対話モデル	システム	ユーザ	混合

表 1: 端末、modality、対話モデルの関係

4 対話実験

2 種類のインタフェースを使用して小規模な対話実験をおこなった．対話人数は 5 名，与えられた対話の目的は，「明日，3 時から 5 時まで第 3 会議室を予約する」というものである．会議室の空き状況は、第 3 会議室しか空いていない状況を人工的に作成、用意した。

対話に使用したインタフェースは携帯電話，および携帯電話と PDA の併用である．2 種類のインタフェースによる対話の違いを探るために，対話の各状態の所要時間を計測し，比較した．対話の状態の分類とその関係を図 6 に示す．

状態 A-B はシステムからのメッセージの出力（音声、文字出力）に要した時間、状態 B-C はシステムからの出力が終了してからユーザーが音声により応答の出力を開始した時間、状態 C-D はユーザが音声による応答の出力を開始してからシステムがその

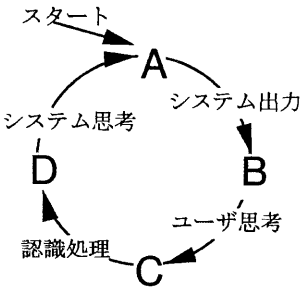

図 6: 対話の状態

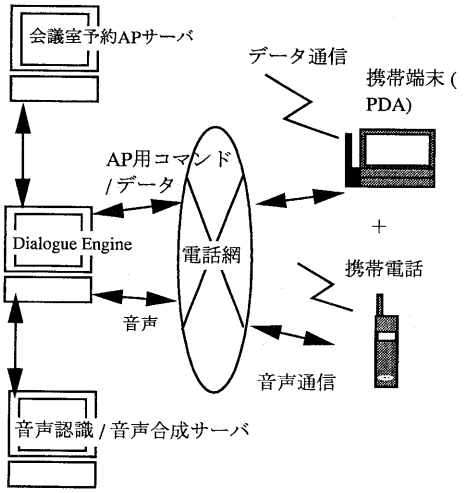

図 7: 作成した対話サービスシステムの構成

応答の認識を完了するまでの時間、状態 D-A はシステムの認識が終了してからシステムの応答が開始するまでの時間である。

実験に使用した対話システムの構成を図 7 に，PDA の画面を図 8 に示す．PHS の宅内モードによるコードレス電話機単独のインタフェース（IF-1）は，完全なシステム主導の音声のみの対話であり，音声入力された情報はかならずシステムによって再確認される．入力音声の始端検出はパワーしきい値によって行なわれる．

コードレス電話および PDA の併用インタフェース（IF-2）では，情報の入力はキー操作と音声のコンビネーションによって行なう。まず、上下カーソルキーにより該当する入力項目を選択し，次にスペースキー

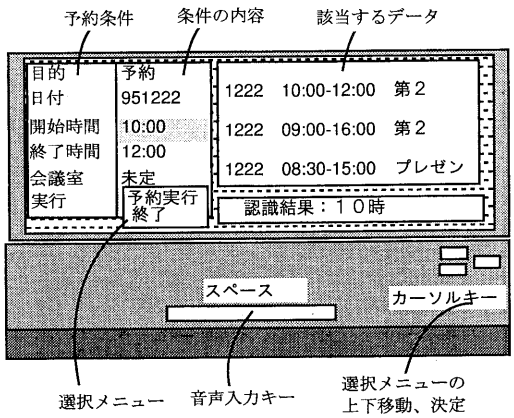

図 8: PDA の画面インタフェース

を押下して音声入力の許可要求を Dialogue Engine に出すと、音声入力の促進合図「どうぞ」がシステムから返ってくる。そして、この合図の後に発声することによって音声入力が行なわれる．促進合図に従って発声することにより、A/D 開始前に音声が入力されてしまうという、複数機器が非同期で動作することによる音声入力の不具合を回避することが出来る。

また、PDA のインタフェースは PDA 単独で使用できるよう設計されており、右カーソルキーによりサブメニューの表示とサブメニュー中のアイテムの選択を行なうことができる。システムの出力は，画面への文字情報出力および音声出力により行われる．

図 7 では電話音声が入力経路として示されているが，今回の実験においては，音声の認識率を高めるために電話回線を通さずにマイク入力の音声を使用した．

2 種類のユーザインタフェースによる対話の各状態での平均所要時間と対話に要した turn 数、トータルの対話所要時間を表 2 に示す．

表からユーザインタフェースの相違による特徴として以下の 3 点があることが分かった．

1. IF-1 の方がシステム応答出力後のユーザからの入力開始が早い．
2. IF-2 の方がシステムからの応答出力の所要時間が少ない．
3. IF-2 の方が少ない対話の turn で目標を達成している．

	携帯電話	併用
A-B	3.2 sec	0.4 sec
B-C	0.3 sec	4.1 sec
C-D	0.5 sec	2.2 sec
D-A	0.7 sec	0.4 sec
所要 turn 数	12	5
平均所要時間	56.4 sec	35.9 sec

表 2: 対話の各状態での平均所要時間

特徴 1 の理由としては、IF-1 がシステム主導の問いかけを行なうため、ユーザが答えるべき内容が分かりやすいこと、システムから出力される音声情報は出力の過程と同時に漸進的にユーザの理解が進むために，出力終了時にはユーザの理解がほぼ終了しているということが考えられる.

IF-2 では電話回線を介して文字情報を表示するのに時間がかかるため、音声入力の開始が遅れることも一因となっている。

特徴 2 の理由は、システムからの応答音声の出力自体に時間がかかるという音声情報特有の性質のためと考えられる。

特徴 3 の理由は、画面表示を利用することにより確認のための対話が不要になったことが大きい。また、IF-2 では混合型の対話モデルの使用により，無駄な対話の turn が省略されたことも影響している。ユーザが予約を希望していた 3 時から 5 時まで空いている会議室が第 3 会議室の一つしかなかったため、システムの方から第 3 会議室の予約を行なうかを聞くというシステム主導の対話がおこなわれ、その結果、会議室名を指定する対話が省略されている。

対話に要した全所要時間は IF-1 よりも IF-2 の方が約 16% 少ない。PDA への予約データや認識結果の表示の遅延を考慮すると、純粋な対話の所要時間の差はより大きいと考えられる。

今回の対話実験では誤認識した turn を削除して所要時間のデータの集計を行なった。誤認識の修正の容易さを考えると、画面上で認識結果が確認でき、任意のタイミングでデータの音声再入力が行なえる IF-2 の方が、ユーザにとって誤認識の修正作業が楽であり、実使用上はさらに対話所要時間の差が開くと予想される。

5 まとめと今後の課題

ユーザの情報環境とサービス内容に応じた対話モダリティの選択と対話戦略を行う対話型サービスサーバ Dialogue Engine を提案し，そのコンセプトに基づいて，ユーザの所有する携帯機器に応じた柔軟なインタフェースをもつ会議室予約オンラインサービスの実験システムを構築した．また，小規模な対話実験を行なった結果，ユーザの使用する端末に応じた対話モダリティと対話モデルの切り替えが有効に動作することを確認した.

タスクの種類やユーザの性質への対話モデルの対応，対話モダリティに応じた応答生成機構の高度化等が今後の課題である.

謝辞

日頃有益な討論をしていただいている情報科学研究所，マルチメディア技術センタ同僚諸氏に感謝します.

参考文献

[1] モバイルＰＣ編，「もっと 200LX！」，Mobile PC,p.38,(June 1996)

[2] 桑野, 石田, 木村, 渡辺, 平岡,「カード型不特定音声認識装置」, 平 7 年度春季音響学会講演論文集,1-Q-33,pp161-162,(1995-3).

[3] 小高、天野、畑岡,「電話回線と LAN を介した音声認識応用の検討」, 信学技報, SP94-55, pp15-21,(1994-11).

[4] 山田、中村、菅野,「映像通信ネットワークのエージェントに関する一考察」, 情報処理学会マルチメディア通信と分散処理研究会資料, 65-15,pp85-90,(1994-5).

[5] 安藤、畑岡,「マルチモーダルなエージェント型ユーザインタフェースの評価と対話制御の検討」, 情報処理学会音声言語情報処理研究会資料, 7-15,pp91-96,(1995-7).

[6] S.Robertson, C. Wharton et al., 'Dual Device User Interface Design: PDAa and Interactive Television', Proc. CHI 96,(1996).