2007－SLP－68　(7)-1
2007／10／19

ProxyAgent：音声認識システム機能拡張のための新たなコンポーネント

小林 哲則† 中野 鐵兵†
†早稲田大学

あらまし:
音声認識システムの機能拡張を，アプリケーションや音声認識エンジンに対して非依存な形で行う枠組みを提案する．この枠組みでは，Proxy-Agent と呼ばれる，アプリケーション・音声認識エンジン・入力デバイスの中間に配備されるソフトウエアコンポーネントを利用する．本稿では，Proxy-Agent の概要と最初のプロトタイプとして開発したソフトウエアについて紹介する．

ProxyAgent: A Software Component For The Speech Recognition System Extension Framework

Tetsunori KOBAYASHI†, Teppei NAKANO†
†Waseda University

Abstract:
This paper presents an extension framework for speech recognition system. This framework is designed to use "Proxy-Agent", a software component located in between applications, speech recognition engines, and devices. By taking advantage of its structural characteristics, Proxy-Agent can provide supplementary services for speech recognition systems as well as user extension. We introduce the first prototype implementation of Proxy-Agent in this paper.

1 はじめに

音声認識システムを利用したアプリケーションにおける，汎用的な機能拡張の要求が存在する．汎用的な機能としては，実環境における利用者の発話内容のモニタリング機能や，モデルデータのネットワーク経由での更新機能，特定の音声インタフェースの利用を可能にするフレームワークの拡張機能等が挙げられる．しかしながら，このような機能を既存の音声認識アプリケーションに対して導入するための共通の枠組みは存在せず，これまでは，アプリケーションやエンジンごとにアドホックな拡張が行われていた．本稿では，アプリケーションや音声認識エンジンに対して横断的に適用可能な汎用的な機能拡張の枠組みを実現するものとして，プロキシエージェントを用いた音声認識システムについて述べる．

2 基本アプローチ

プロキシエージェントとは，アプリケーションプログラム，音声認識エンジン，入力デバイス，の間に配備されたソフトウエアコンポーネントであり，それらの協調動作を担う (図 1)．プロキシエージェントは，アプリケーション／エンジン依存の機能拡張と同時に，アプリケーション／エンジン横断の機能拡張を可能にする．この枠組みでは，アプリケーションは音声認識エンジンを Engine-Adapter によって提供されるインタフェースを通して操作する．Engine-Adapter は，音声認識エンジンのプログラムを含む，1 つ以上のプラグイン群から構成される仮想音声認識エンジンである．Enigne-Adapter は，その入力データを Proxy-Agent が管理する Device-Adapter 経由で取得する．Device-Adapter は，実際の入力デバイスからのデータ取得ロジックを包含したデータ提供オブジェクトであり，実際に処理の対象となるデータを取得する．

この枠組みにより，アプリケーションや音声認識エンジンの主とする機能ではないが重要な機能の埋め込みを容易にする．アプリケーションやエンジンに直接機能の埋め込みを行う枠組みと比較すると，開発者はこれを，開発するアプリケーションのドメインや利用するエンジンに依存しない，音声認識アプリケーションの共通の開発基盤として利用することが出来る．これにより，広範囲のアプリケーションでの拡張機能の適用が可能となる．以下に，Proxy-Agnet によって実現を目指す，音声認識システムの標準的な機能拡張のための枠組みが備える主要の機能を示す．

Extension capability 全ての拡張機能が Proxy-Agent に対するプラグインとして拡張可能であること

Networking capability ネットワーク経由でのサーバ連携が可能であること

Monitoring capability 実利用環境における実行時の情報が取得可能であること

Upgrade capability 実利用環境におけるプラグインの継続的な更新が可能であること

Sharing capability 言語リソースやコンポーネントの実装，フレームワーク等の開発者間での共有が可能であること

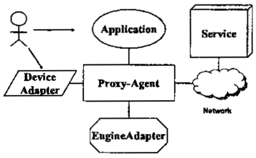

図 1: Proxy-Agent Overview

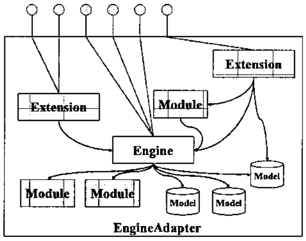

図 2: Engine-Adapter Overview: Interface of Engine-Adapter is defined as an aggregation of plugins interfaces

3 仮想音声認識エンジン

Engine-Adapter は，プラグインによる機能拡張の枠組みを有する仮想音声認識エンジンとして提供される．プラグインは，音響モデルや発話区間検出機能モジュール等の音声認識エンジンの構成要素から，音声認識エンジンそのもの，またはアプリケーション開発者が用意したアプリケーション特有の音声認識エンジンの操作プラグインまで自由な粒度で構成される．そのインタフェースは，音声認識エンジンと拡張機能のインタフェースの集合として公開される (図 2)．操作プラグインは，従来はアプリケーション側に記述されていた，音声認識エンジンの操作ロジックを埋め込んだプラグインであり，アプリケーション毎に用意される．音声認識エンジンと蜜に結合した操作プラグインが Engine-Adapter に登録されることにより，それを呼び出すだけのアプリケーションと音声認識エンジンとの関係を疎結合とすることが可能となる．また，さらにそこで作成したプラグインを他のアプリケーションからの利用することで，音声認識アプリケーションに関する知見の共有を可能にする．

また，Engine-Adapter は，デバイスからの入力の取り扱いや，コンポーネントのプラグイン，ロードの枠組みに関する基本機能のみを提供し，標準インタフェースの定義等はエンジン開発者が自由に行えるようにする．すなわち，エンジンの基本構成や部品間のインタフェースは，音声認識エンジンプログラムごとに独自の観点で定義する．しかしながら同時に，全ての部品が同一のプラグインの枠組みで開発されるために，VAD や雑音除去といった一部のプラグインを組み込むことも可能である．これによりエンジンの独自性を妨げることなく，エンジン横断的な部品から，あるエンジン専用の部品まで，様々な部品を組み合わせての音声認識エンジンの構築が可能となる．

4 プロトタイプ開発

プラットフォームとして Eclipse RCP(Rich Client Platform)[3] を採用し，設計・開発を行った．Eclipse RCP とは，非常に拡張性の高いプラグインアーキテクチャを提供するプラットフォームであり，Java ベースのオープンソース統合開発環境である Eclipse の基盤として開発されている．本研究ではプロキシエージェントを Eclipse プラグインとして開発し，RCP Application として実行可能にした．つまり，前節で説明した Proxy-Agent, Engine-Adapter, Device-Adapter を全て Eclipse Plugin として実装した．また，それらに対するプラグインも全て Eclipse プラグインとして実装可能にした．これにより，Eclipse IDE の提供する統合開発環境を利用した Proxy-Agent のプラグイン開発，Eclipse が利用する標準的な枠組みに従ったネットワーク経由でのプラグインの管理・更新機能, 及び Java によるマルチプラットフォーム性等をサポートした．

また，このプロトタイプとして開発した Proxy-Agnet を用いて，データ入力アプリケーション [1], アプリケーションコントローラ [2] 等の開発を行った．音声認識エンジンとしては，Java ベースのオープンソース音声認識エンジンである CMU Sphinx4[4] と，Java 経由での呼び出しを可能にした Julian[5] と VORERO[6] を利用可能にした．Sphinx4 は，その構成要素であるモデルやコンポーネントを Eclipse Plugin として登録出来るように拡張し，プラグインの組み合わせによる Engine-Adapter の作成を可能にした．これらの開発を通して，Proxy-Agent によって実現されるモニタリング機能と，プラグインによる音声認識システム・フレームワークの拡張による効果の有効性を確認した．

5 むすび

本稿では，Proxy-Agent と呼ばれるソフトウエアコンポーネントを利用した，音声認識システムに対する汎用的な機能拡張の枠組みに関する提案を行った．また，その概要について述べ，最初のプロトタイプとして開発したソフトウエアの紹介を行った．

参考文献

[1] Teppei NAKANO, et al., “Efficient Monitoring Capabilities of User Behavior for Speech Recognition Applications Based on Proxy-Agent Architecture,” 2007-SLP-65, pp.23–28, IPSJ SIG, 2007.

[2] Tomoyuki KUMAI, et al., “A Proposal and Evaluation of New Speech User Interface Based on Functional-Structure,” 2007-SLP-67, pp.109–114, IPSJ SIG, 2007.

[3] `http://www.eclipse.org/rcp/`

[4] `http://cmusphinx.sourceforge.net/sphinx4/`

[5] `http://julius.sourceforge.jp/`

[6] `www.asahi-kasei.co.jp/vorero/`

自動車運転行動中発話の分析

旭化成株式会社　音声ソリューションビジネス推進部
庄境 誠, 加藤 智之, 岡本 淳

アブストラクト

カーナビなどの車載情報機器のハンズフリー音声インタフェースの搭載率は高まっているが，利用率が高いとはいえない。利用率に相関の高い実環境性能すら，明らかになっていないのが実情である。そこで，260 名分の自動車運転中発話コーパスを新たに収集し，低認識性能の被験者に注目して行った物理量と認識性能との相関分析結果について論じる。

Analysis of Utterance during Action of Driving Car

Speech Solutions, Asahi Kasei Corporation
Makoto Shozakai, Tomoyuki Kato and Jun Okamoto

Abstract

Although a loading rate of hands-free voice user interface capability into in-vehicle information appliances such as GPS navigation system has become high, a usage rate of the capability is not necessarily high. Even performance in adverse environments, which is highly correlated to the usage rate, is not revealed yet. We collected the speech corpus of utterance of 260 persons during action of driving car. Results of correlation analysis between acoustic parameters and performances for low performance subjects are discussed.

1. 背 景

音声認識のアプリケーションの中で，ハンズフリーインタフェースの必要性が明確な車載情報機器（カーナビ，カーオーディオ，携帯電話）への音声インタフェースの搭載率は，他の潜在的アプリケーションに対して高いと言える。しかしながら，それらの機器のハンズフリー音声インタフェースの使用率は必ずしも高いとは言えないのが実情である。その原因についても，様々な議論がなされてきた。

万人に使用される車載情報機器のインタフェースとして期待される役割を考えれば、様々な未解決課題の中でも，実環境における十分高い対雑音性能，対話者性能，対語彙性能の確保の問題を無視するわけにはいかない。本論では，この問題について，議論する。

2. 認識性能の表現

実環境における対雑音性能，対話者性能，対語彙性能には分布（ばらつき）が存在する。例えば，認識性能に関して降順に話者（利用者，被験者）を横軸に並べた場合，縦軸の認識性能は図 1 のような形状を取る場合が一般的である。ここで，A 点(*, y)は，全話者の平均性能を表す。B 点(0, y)は，全話者の最高認識性能を表す。W 点(100, y)は全話者の最低認識性能を表す。I 点(x, y)は，認識性能分布の屈曲(Inflexion)点を表す。認識タスクにも依るが，B 点から I 点まではなだらかに認識性能が低下し，I 点から W 点までは急激に認識性能が低下する傾向が一般的である。

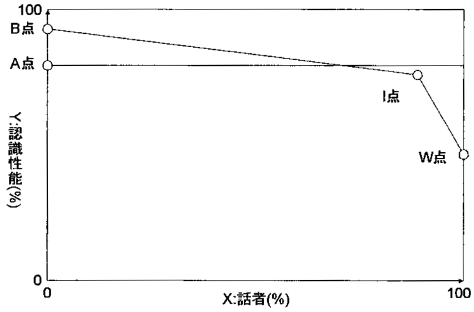

図 1 認識性能分布の概形

従来，認識性能は平均認識性能である A 点の y 値のみで表現されることが多かった。大語彙連続音声認識のような難しい認識タスクの場合も，A 点の 100-y 値（誤り率）がどのくらい改善したという研究が多い。その一方で，W 点や I 点の周辺に着目し，それらの y 値が 100 から乖離している原因の本質に迫る研究の取り組みがあまり見られないのが実情である。

一方，音声認識機能を実用化の場面で使う側の立場で，音声認識技術を外の世界から眺めた場合に，まず関心があるのは，W

点のｙ値である。この値は，実環境での認識性能の品質保証（最低保証）に直結するからである。しかし，実環境においては，W点のｙ値が90(%)を大きく下回る場合も多い。その場合でも，I点の(x,y)値に着目し，それらが(80,90)，(90,90)，(95,95)などの値以上であることが確認できれば，製品・サービスの企画コンセプト次第では実用化に踏み切るという判断もあろう。

3. 自動車運転行動中発話コーパスの収集

自動車運転行動中の発話における対雑音性能，対話者性能，対語彙性能を明らかにするためには，それらの分析が可能な発話コーパスの確保が必要である。自動車運転行動中の発話コーパスとしては，CIAIRコーパス[1]が知られているが，被験者の発話内容が完全に制御されておらず，対雑音性能，対話者性能，対語彙性能を客観的に分析するには好適ではなかった。そこで，以下の条件を満足する自動車運転行動中の発話コーパス（260名）を新たに収集した。

1）録音ゲインが全収録で固定
2）発話内容が複数の語彙種を含んでおり，全話者で共通
3）発話の際の自動車運転行動が，全話者で共通

これにより，雑音性，話者性，語彙性を切り分けて分析できる。収集コーパスの概要を表1に，収録条件を表2に示す。

表1 収集コーパスの概要

被験者種別	走行状態	人数	目的
一般被験者（一般人）	アイドリング/コース内走行	男女各110名	日本語発声の話者による広がりを把握
プロ被験者（運転のプロ）	アイドリング/市街地走行/高速走行	男女各20名	より現実的な運転行動中音声を把握

表2 収録条件

標本化周波数	ビット長	チャネル数	被験者への語彙の指示	使用車両
48kHz	24bit	4ch	TTS	Vitz(TOYOTA)

4. 発話コーパスの物理量分析

収集した発話コーパスの対話者性能を解析するために，以下の物理量の統計分析を行った[2]。

(1)音圧：SS後の10msフレーム最大平均パワー
(2)SNNR：プリエンファシス後の発話/雑音区間のパワー比(dB)
(3)発声速度：単位時間当たりのモーラ数(値が大きいと早口)
(4)滑舌度：特定話者音響モデルの5母音間距離(値が小さいと滑舌が悪いといえる)
(5)話者固有度：不特定話者音響モデルに対する特定話者音響モデルの距離(値が大きいとスペクトルの固有度が大きく，特徴的な声といえる)

5. 低認識率被験者の物理量と認識性能の相関分析

表3の認識タスク条件で求めた認識性能の分布を図2に示す。図2のW点，I点付近の丸で囲まれた2名の女性プロ被験者の物理量のヒストグラムを全女性プロ被験者の物理量のヒストグラムの上に重畳表示した結果を図3に示す。2名の被験者の特徴的な物理量を楕円で囲んで示す。SNNR，滑舌度，話者固有度が低認識性能に強い相関を持つことが示唆される。

表3 認識タスク条件

音響モデル	Monophone/43音素/単一正規分布	Leave-one-outによる女性不特定話者モデル
待ち受け語彙	約1000単語	カーナビ用想定コマンド
使用音声	女性プロ被験者(20名)/高速走行中音声/マップランプ付近設置マイク収録音声(1ch)	
特徴量	MFCC(12)，ΔMFCC(12)，ΔlogPower(1)	サンプリング周波数16kHz

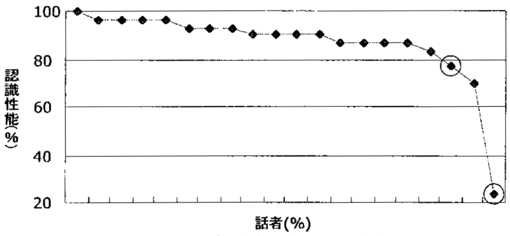

図2 女性プロ被験者の認識性能分布

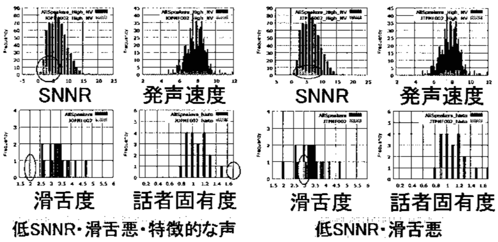

図3 W点(左)，I点付近(右)の被験者の物理量分析

6. 今後の予定

低性能や誤認識の原因は多様である。上述の相関分析を詳細に進め，対雑音性能，対話者性能，対語彙性能の低下に対して，より大きな影響を与える原因を特定し，順次，対策を講じることにより，I点の(x,y)値が(100,100)に近づく効果を定量化して行く予定である。

参考文献

[1] 河口，松原，岩，梶田，武田，板倉：実走行車内における音声データベースの構築，情報処理学会，音声言語情報処理研究会，SLP-30-12，2000.

[2] 加藤，岡本，庄境：自動車運転行動中発話の日本語音声コーパスの物理量と認識性能の相関分析，日本音響学会，音講論，3-Q-25，2007.9.

UI 設計とユーザビリティ
—音声インタフェースの課題—

石川 泰
三菱電機株式会社 情報技術総合研究所
Ishikawa.yasushi@eb.MitsubishiElectric.co.jp

あらまし 音声認識技術の発展により、音声インタフェースの実用化が進んでいるが、その利用分野の拡大、普及は十分に進んでいるとは言えない。本稿では、音声インタフェースをタスク達成のインタラクションの視点で、GUI と比較しながら、ユーザビリティ視点で整理し、その問題点を検討することで、今後の音声インタフェース設計の課題を議論する。

キーワード 音声認識 音声インタフェース 音声対話 GUI UI 設計 ユーザビリティ

UI Design and Usability
- Subjects of Speech Interface -

Yasushi Ishikawa
Information Technology R&D Center, MITSUBISHI Electric Corporation

Abstract In association with technical progress of speech recognition, products with speech interface have been developed. However, the spread of application areas and the diffusion are still inadequate. In this paper, the usability of speech interface is described from the viewpoint of interaction for task completion comparing GUI. And from investigation of problems, the subjects of speech interface design are discussed.

Keywords Speech Recognition, Speech Interface, Speech Dialog, GUI, UI design, Usability

1. はじめに

音声認識技術は、1990 年代に入り、統計的手法の導入により不特定話者音声認識が十分に高い性能を示すようになり、プロセッサ技術の進歩により、製品開発が本格化した。電話系音声認識技術による CTI (Computer Telephony Integration) システム、コールセンター応用が注目され、ディクテーション S/W が販売され、携帯電話などへの利用が図られてきた。しかし、現在、カーナビゲーションでの利用がハンズフリー、アイズフリーという利点から実用システムとして継続的に注目され、利用されてはいるものの、音声認識技術者の期待からは、その利用分野も、普及も十分とは言えない。

一方、ユーザインタフェースの重要性は、機器・システムの機能の複雑化、扱う情報量の飛躍的拡大、携帯電話やディジタル音楽プレーヤなど小型機器の普及を背景に、製品価値を決定する大きな競争軸としてますます高く認識されており、高度なグラフィック処理による GUI が開発されている。音声認識の利用を推進す

るためには、単に音声が人間にとって最も自然な情報伝達の手段であるということに基づいた音声認識の有効性の主張や、ハンズフリー・アイズフリーといった制約からの利点の訴求ではなく、音声認識の特長、制約・問題点を考慮したUI設計技術の検討が必須である。

もちろん、これまでも、音声認識の実用化を阻害する要因の分析や、課題の検討が行われ、すぐれた知見が示されてきた[1][2][3]。音声認識技術の問題、音声対話の制御技術・対話記述の技術の問題、さらには、人間機械間のインタラクションがどうあるべきか、これら議論はすぐれた先行研究にゆだねることとし、本稿では、GUI との比較により、音声インタフェースがどのような特徴を有するかを検討する。

2 章では、ユーザインタフェースにおけるユーザビリティについて概説し、3 章では、従来の音声インタフェース設計例を GUI と比較することで、ユーザビリティ上の問題の明確化を試み、4 章において、これらの検討から音声インタフェースの課題を議論する。

2. UI とユーザビリティ

ユーザビリティはNielsenによれば、「学習容易性」「効率性」「記憶性」「エラー」「満足度」により決まるシステムの有用性と定義される[4]。この定義は、人間工学、認知工学、感性工学の観点から考察されたものであり、UI(ユーザインタフェース)の問題点抽出の視点で重要な着眼点である。しかし、評価の観点では、相互の関係性も強いため、1998 年に制定された ISO9241-11 では、「指定された利用者によって、指定された利用の状況下で、指定された目標を達成するために用いられる際の有効さ、効率および、満足度の度合い」として定義されている。ここでは、この定義に基づき、主に有効さと効率について検討する。ここで、有効さは、「目標を達成する上での正確さと完全性」を、効率とは「ユーザが目標を達成する際に、正確さと完全性に費やした資源」と定義される。ただし、これらを一般的に論じることは極めて困難であり、また音声認識における誤認識の問題を含めると、複雑になるため、ここでは、タスクを単純化した場合についての GUI との比較により検討する。

また、ユーザビリティの評価においては、しばしば、「目標の達成」、すなわち評価タスクを規定した評価が行われる。しかし、携帯電話やカーナビゲーション、ディジタルテレビなど、今日の情報機器では、その機能の高度化・多様化により、ユーザがシステムに対して使用時に抱く目標が、対象システムで実現できるかどうかがわからない、あるいは逆に、システムが保有する機能がユーザに利用されない、つまり目標とされないことが頻発する。すなわち、実際のシステムとユーザ期待とのミスマッチである。UI デザインでは、このようなミスマッチを防ぎ、すべての機能をユーザに提供することが重要となる。広くは、「有効さ」「効率」にかかわる問題であるが、具体的なミスマッチがどのように起きるかを含め検討することとする。

3. GUI と音声インタフェース

対象システムの機能、UI を一般化することは困難であるが、ここでは、単純化し、機能選択型のタスクと、データ入力型のタスクに分類する。さらに、単純な音声のみのインタフェースと、GUI の典型的手法について比較検討する。

3.1. 機能選択タスク

3.1.1. GUI

GUI に着目すると、ユーザが希望する機能を多くの機能から選択する UI は、図 1 のように、木構造で表現される階層化されたメニューの選択で表すことができる。すなわち、ある状

況からのユーザの目標に相当するノードへの到達問題である。

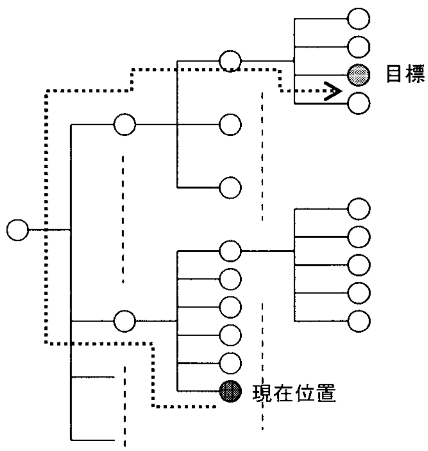

図 1. GUI による機能選択タスク

したがって、効率は、現在の位置から、目標のノードまでの、操作回数(階層の深さ)と、各階層でノードを選択するために要する資源(労力)によって決まる。一般に、このノードの選択のためには、言語表現によるメニュー、あるいは、グラフィカルアイコンによる選択、またはその組み合わせが用いられる。労力は、それらのメニューの理解、選択のための操作に費やされる。

一方、「有効さ」は、提示される用語や、グラフィカルアイコンがユーザの目標、あるいはその上位概念をあいまい性なく、表現しているかによって、正しく目標に到達できるかという問題として考えられる。

3.1.2. 音声インタフェース

機能選択タスクの音声による実現手段は種々考えられる。ひとつには、GUI と完全に一致した処理体系での実現であり、ポインティングデバイスあるいは、H/W スイッチの変わりに音声を利用しようとするものである。このような設計思想であれば、基本的に比較は、音声入力の労力と GUI、H/W 操作の労力との差異、および、グラフィカルアイコンあるいは、表示文字と、音声コマンドの想起の比較となる。しかし、一般には、音声認識の語彙の大きさの活用が図られる。その典型例がすべての状態から、目標操作に到着するための音声コマンドを定義する手法である。

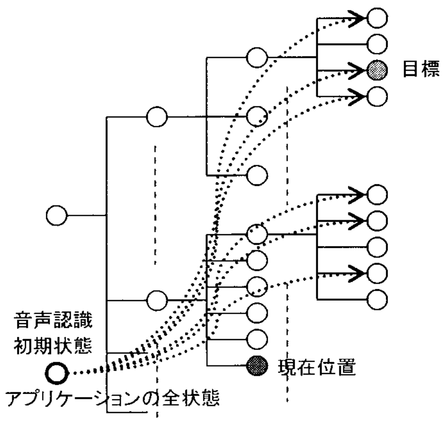

図 2. 音声コマンドによる機能選択タスク

このような方法をとることにより、操作回数が減少し、効率向上が期待できる。しかし、「有効さ」について考えると、目標機能を表現する、設計者の発想による一意な「音声コマンド」をユーザが選択できるかどうかという問題が生じる。音声認識では、音声コマンドが「日常使用している言葉」であるという、人間同士のコミュニケーションのアナロジーから、容易にユーザが記憶できると誤解される場合がある。機器の機能が複雑になればなるほど、すなわち、GUI 操作のインタラクションが増え、1 発話による入力の有効性が高い場合ほど、このコマンドの一意性の保持と、ユーザの期待との合致は困難になり、音声インタフェースの有効さを著しく低下させる要因となる。

連続音声認識、自然言語理解技術は、ユーザの目標に対する音声操作の自由度を高めるものではあるが、自由度を高めることは、ユーザ発話が、目標のノードだけではなく、GUI におけるその上位ノードの表現である場合や、それが現在の状態に依存した表現になることをも許容しなくてはならず、困難な課題であることは言うまでもない。

さらには、GUI の操作が、操作の範囲をメニュー、アイコンなどでユーザに提示しているのに対して、音声インタフェースでは、一般に操作の範囲が明示されない。したがって、上述のように、ユーザがシステム機能について理解していない場合、実際にない機能の要求がなされ、また、GUI 操作により、機能の学習ができる、すなわち、提示されることにより、ユーザが知らない機能を学習できるのに対し、音声インタフェースのみからは、学習が進まない可能性も高い。

3.2. データ入力型タスク

一般にスロットフィリングで表現されるタスクを考える。スロットに入る値の種類が少ない場合はメニューの選択としての実現ができるが、ここでは、スロットに入力する値の種類が多い、あるいは、制限がない場合を考える。

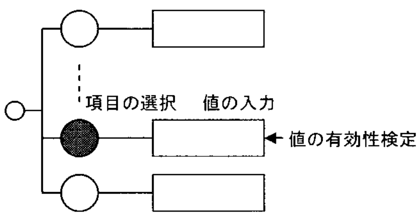

図 3. データ入力型タスク

このタスクを一般化すれば、図 3 に示すように、複数項目についての値入力のタスクとなる。したがって、ユーザは、なんらかの方法で、項目を選択し、値を入力、さらには、その値がシステムにとって意味のある値かどうかが検定されるとともに、用意された項目の中からタスクを完了するために必要な項目をすべて入力することで、タスクが完了する。

3.2.1. GUI

GUI における効率は、いかに容易に項目が選択できるか、そして値の入力の労力がどのようなものかにより決まる。項目の選択では、ポインティングデバイスなどを用いてユーザが選択するか、システムが順番を規定するかいずれかの方法により実現され、その効率をタスクの特異性にあわせて検討すればよい。値に入力については、値の種類が比較的少ない場合には、GUI のリストセレクションボックスや、ラジオボタンによる選択が有効であり、この場合、値の有効性の検定も同時に可能となるというメリットがある。値の種類が大きい場合は、キーボードなどの入力手段を用いて入力することになり、労力が大きいことが予想できる。

また、「有効さ」については、項目の選択のために、表示されたテキストあるいはグラフィックアイコンがユーザの期待といかに合致し、誤らずに選択できるか、システムが項目の入力順序を規定しているとき、その順序がユーザの期待に合致し、判断誤りを生じさせないかが検討項目である。また、ユーザの入力がシステムの入力値として有効なものとして正しく受け入れられるか、さらには、ユーザが入力したい項目と、システムのタスク完了のために必要な項目とに不一致が生じないかどうかが重要な要素となる。

3.2.2. 音声インタフェース

音声インタフェースの実現では、項目選択の実現手段として、GUI 同様システムが入力項

目の順序を規定する場合と、ユーザが項目を選択できる場合とが考えられる。これは、基本的に上記のとおり音声インタフェース特有の問題ではない。しかし、以下の点でそのユーザビリティには大きな相違を生じる可能性がある。すなわち、GUI が順番を規定していても、入力のタイミングまでは規定しないのに対し、音声対話においては、システムが入力を要求する場合、入力のタイミングまでを規定し、かつ、その場合に、ユーザに要求どおり発話する以外のことができない状況を強いる場合が多い。このことは、ユーザに大きな労力を要求することになり、とくに、その順番がユーザの期待と一致しないときは大きな問題になる。また、ユーザが選べる場合には、項目をあらわす語と設計者の期待との一致が必要であり、これが満たされず、ユーザの期待と異なる項目が選択された場合、著しい効率の低下、あるいは、有効さの低下につながる。

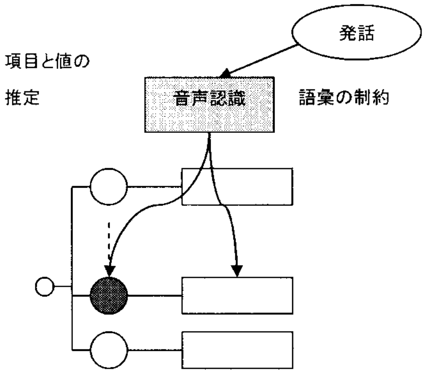

図 4. 音声インタフェースによるデータ入力

また、値の入力においては、ユーザの発話が、正しく認識できれば、効率の上で、他の入力手段に比べ有効である可能性は高い。しかし、「有効さ」について考えると、システムの期待する入力値として有効かどうかが、入力時点では判断できず、また、リジェクトが正しく行われない限りは、システムは、入力値を有効な値にマッピングする(誤認識)するという問題が生じる。このとき、何らかの方法で、ユーザが入力した内容が有効ではなかった、つまり語彙外発話であったことが判断できれば、有効さについても、大きな問題を生じることはないが、音声認識はたとえ、待ち受けている語彙を発声しても誤る場合があることを理解しているユーザには、語彙外か誤認識の判断はできない。

このことが、有効さの大きな低下につながる。

3.3. タスクに依存しない問題

一方、GUI と音声インタフェースの比較では、現状では GUI がはるかに入力の多様性がある点を見逃してはいけない。すなわち、タッチパネルに触ること、キーを押すこと、ダイアルを回すこと、などに加え、動作感覚の制御に基づくハプティックデバイスや、タッチパネルなどの上や、カメラを用いたモーションの認識など、近年急速に多様な入力手段が実現しつつある。これらは、一見発話による明確な意思の伝達に比べ、あいまいな感があるが、実際には、現実のものの操作とのアナロジーにより、直感的に操作の質が理解しやすいメリット、さらには、新たな操作方法についての驚き、楽しさなどこれまでに議論してきた「有効さ」「効率」にとどまらず、「満足」への訴求要因となっている。

4. 音声インタフェースの課題

4.1. 「効率」と「有効さ」

音声インタフェースの検討は前述のように、種々の側面からすぐれた検討がなされてはいるが、実際の製品開発の現場では、インタフェースのユーザビリティ視点からの検討が十分には行われず、「有効さ」における問題を、場合によっては「言語表現であること」で解決できているかのように考え、音声認識の「効率」に

ついての基本的なポテンシャルを訴求するUIデザインが行われることが多い。

これが、正しくないのは、従来のコンピュータのインタフェースにおける「コマンド」の入力手段が、キーボードから、音声認識に変更されただけの状況を考えればあきらかである。

筆者は、一般的な解決方法を提示はできないが、以下の課題を検討していくことで、知見を蓄積し、音声認識の有効性をそれぞれの対象で効果的に活用する手法が製品として広がり、音声認識技術の広い実用化が果たせるものと考える。

1） 複雑なインタラクティブシステムの評価手法の検討と標準化

ともすると、タスクの理解があるものとして評価がなされ、現状の組み込みシステムや、ディジタル機器のインタフェースに見られるシステムの機能とユーザ理解のミスマッチが評価されない。評価方法を確立し、評価と改良のプロセスが UI デザインの場で行えるようにする必要がある。

2） 音声の「有効さ」の検討

音声インタフェースは、ユーザにとり理解が困難な現時点で操作可能な機能と、その表現方法が設計者理論で規定されてしまっている。このミスマッチを防ぐためには、単なる効率の訴求によるデザインをせず、生じえるミスマッチのインスペクション評価を十分に実施することが必要である。また、機能の自己表現性は、GUI がすぐれているとすれば、GUI との一貫性や、あいまいな入力を設計レベルで許さないのではなく、あいまい性を解消し、学習できる音声インタフェース設計も重要である。

3) GUI で困難な問題の検討

上述のように GUI は自己表現性が高いが、できることの制約や、深い階層の機能選択タスクではその構造の理解ができないための「有効さ」の低減が生じる。音声で直ちにこれを解決するのではなく、音声入力によるヘルプ機能など、相補的なインタフェースの設計も重要である。

4） 音声インタフェースの「満足」

GUI は直感的な操作手段を数多く、提供し、提示するコンテンツ・システム状態のグラフィカルにリッチさでユーザの「満足」の向上を図っている。データ入力や機能選択といった明確な目的達成場面ではない、利用方法、他の入力手段の補助手段としてリッチなインタフェースを実現する方法、あるいは、タスクの検討も必要である。

5. まとめ

音声インタフェースとGUIを比較し、インタフェースのユーザビリティを考察することで、課題を検討した。具体的な解決策の提示は筆者には困難であり、検討も稚拙ではあるが、今後、音声開発者の相互議論を発展させ、音声インタフェース実用化を推進したい。

【参考文献】

[1] 嵯峨山、「音声認識の実用化を阻むもの」信学ソサイエティ大会, Vol.1995 情報・システム,pp.312-313 (1995)

[2] 畑岡、「音声技術実用化の課題と取り組み」、情処学会、研究報告 SPL-055 (2004)

[3] 庄境ほか、「音声言語情報処理技術デベロッパーズフォーラム」情処学会、研究報告、SLP107 (2006)

[4] ニールセン、「ユーザビリティエンジニアリング原論」東京電機大学出版局 (2002)

小型音声対話モジュールの開発

佐藤 幹　岩沢 透　杉山 昭彦　高野 陽介　*藤田 善弘

NEC 共通基盤ソフトウェア研究所, *NEC 企業ソリューション企画本部

小型音声対話モジュールの開発について報告する。本モジュールは、実環境における音声対話に必要な機能として、マイクの指向性を人の発話した方向に向けるための音源方向検出機能、マイク入力信号に混入する周囲雑音成分の消去を行う雑音消去機能、入力音声を認識するための音声認識機能、認識結果等をユーザに伝えるための音声合成機能を備える。NEC において研究開発中のパーソナルロボット PaPeRo に搭載されるこれらの機能を、携帯電話用アプリケーションプロセッサを利用することにより、様々な機器への組み込みを容易とする小型モジュール上で実現している。

Implementation of a Compact Speech Dialogue Module

Miki SATO, Toru IWASAWA, Akihiko SUGIYAMA, Yosuke TAKANO, *Yoshihiro FUJITA

NEC Common Platform Software Research Laboratories, *NEC Enterprise Solutions Planning Division

This paper presents development of a compact speech dialogue module. This module consists of direction of arrival estimation (DOA), noise cancellation (NC), speech recognition (SR), and text-to-speech conversion (TTS). DOA is essential to adjust the microphone direction toward the speaker. NC is useful to reduce undesirable influence by ambient noise and interference. SR recognizes the input speech and TTS conveys the recognition result to the user. These functions, equipped with in PaPeRo, are implemented on an application processor that was developed primarily for mobile phones.

1. はじめに

近年、ロボットやカーナビ等の様々な機器のインターフェースとして、音声対話機能が注目されている。実環境でスムーズな音声対話を実現するために、様々な雑音が存在する環境に対処できる耐雑音性能が求められる。耐雑音性能を向上させるために、指向性マイクが広く用いられる。マイクの指向性を音声の到来方向に一致させるために、音源方向を推定することが重要である[1]。マイクの指向性だけで抑圧できない雑音や妨害信号は、雑音消去処理によって影響を軽減することが有効である[2]。

一方、実環境での音声対話の困難さは、音声対話機能を搭載する端末の開発にとって障害となっている。そこで、実環境での音声対話に必要な機能を小型・低消費電力・低コストの組込用モジュールとして提供することにより、開発の困難度を低減することができる。

我々は、パーソナルロボット PaPeRo[3]を用いて、実環境を対象とした音声対話技術の研究開発を行っている。この技術を踏まえ、かつ小型・低消費電力・低コストを達成するために、携帯電話用アプリケーションプロセッサを利用して、PaPeRo における音声対話機能と同等の機能を搭載する音声対話モジュールの開発を行った。本稿では、音声対話モジュールの機能、ハードウェア、及びその実装形態について報告する。

2. 音声対話モジュールの機能

音声対話モジュールの全体構成を図 1 に示す。本モジュールは、実環境における音声対話の基本機能として、音源方向検出機能、雑音消去機能、音声認識機能、音声合成機能を有する。これらの機能は、RT (robot technology) ミドルウェアに対応した RT コンポーネントとして動作し[4]、ネットワーク接続された外部アプリケーションから利用することが可能である。

音源方向検出機能は、3 または 4 つのマイクの入力信号を用いて、音源方向の推定を行う。本機能は、近接音場を想定した方向推定[1]を行うことにより、音源とマイクが同一平面上になく十分に距離が離れていない場合も、高精度な方向検出を可能とする。

雑音消去機能は、音声用及び雑音用の 2 つのマイクの入力信号を用いて、音声側入力信号に混入する雑音成分を推定することにより、雑音の消去を行う。本機能は、雑音側入力信号に混入する音声成分の推定も行うことにより、低歪の雑音消去が可能である[2]。

音声認識機能は、音声認識辞書に登録した単語を対象とした音声認識処理を行う。音声認識辞書は、SRGS(Speech Recognition Grammar Specification)[5]準拠のXML形式で記述される。本機能は、不要雑音を認識単語と誤認識させないためのリジェクション機能[6]も備える。

音声合成機能は、入力テキストの読上げを行う。本機能は、組込用途向けのコンパクトかつ高音質な音声合成エンジンにより、合成発話音声の生成を行う。

3. 音声対話モジュールのハードウェア

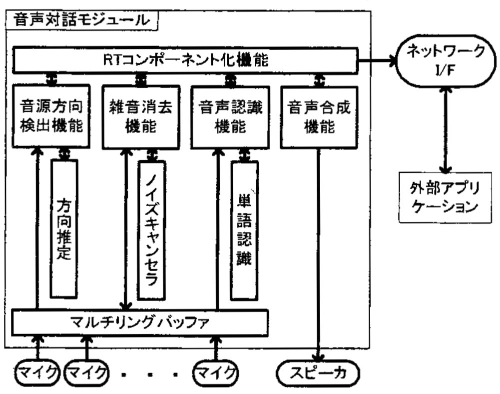

図 1 音声対話モジュールの全体構成

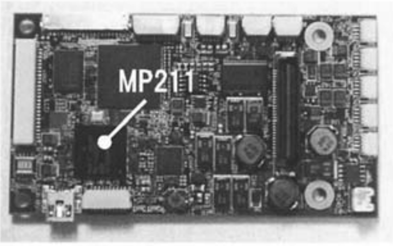

図 2 音声対話モジュールの外観

表 1 音声対話モジュールの仕様

項目	仕様
CPU	ARM9(192MHz) ×3, DSP(SPXK6 192MHz)×1
メモリ	128MB, +128MB (拡張ボード使用時) フラッシュ 64MB
音声 I/F	マイク入力 2ch ×2 系統 +16ch (AUDIO ボード使用時) スピーカ出力 2ch
画像 I/F	カメラ入力×2 系統、ビデオ出力, LCD 出力
その他 I/F	USB, LAN, IrDA, GPIO CF Card (拡張ボード使用時)
サイズ	55mm × 100mm × 32mm (AUDIO ボード、拡張ボード使用時)

音声対話モジュールの外観を図 2 に、仕様を表 1 に示す。本モジュールは、NEC エレクトロニクス社製携帯電話用アプリケーションプロセッサ MP211 (ARM9×3, DSP×1 で構成) を搭載し、Linux OS によって動作する。音声入力インターフェースは、メインボード上に 2 チャンネル同期マイク入力 2 系統を備え、拡張用 AUDIO ボードにより、16 チャンネル同期マイク入力を増設することが可能である。その他には、スピーカ出力×2、カメラ入力×2、LCD 出力、USB、LAN などの周辺インターフェースを備え、拡張ボードにより、メモリカードの使用が可能である。

4. 基本機能の実装形態

音声対話モジュールの基本機能は、MP211 上の ARM9(192MHz) 1 個と DSP(192MHz)1 個により実現している。ARM-DSP 間の基本機能分割を図 3 に示す。音声認識機能、音声合成機能、RT コンポーネント化機能は ARM 上で動作する。音源方向検出機能は、ARM 上で動作する音源方向検出制御部と、DSP 上で動作する方向推定コア部から構成され、3 または 4 マイクの入力に対応できる。雑音消去機能は、ARM 上で動作する雑音消去制御部と、DSP 上で動作するノイズキャンセラコア部から構成される。

マイクに入力された信号は、11025kHz でサンプリングされ、ARM 用メモリに配置されたマルチリングバッファに書き込まれる。音源方向検出制御部は、外部から動作開始命令を受けると、マイク数(3 または 4)に対応した入力信号をマルチリングバッファから取り出し、ARM-DSP 間の共有メモリにコピーする。方向推定コア部は、入力信号を共有メモリから取得し、FFT、白色化、符号化を行い、マイクペア毎の相互相関から推定方向を求める。マイクペア数に対応した相関値は、共有メモリを介して ARM に伝達される。音源方向検出制御部は、相関値で表された複数の方向推定値を個別に最適化した後、最終方向決定処理[1]によって音源方向を決定する。方向推定コア部の方向推定値計算は 1024 サンプル毎、総演算量は 35MIPS である。

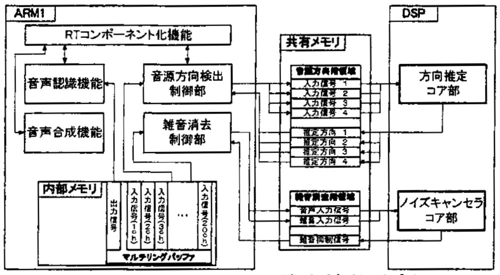

図 3 ARM-DSP 間の基本機能分割

雑音消去制御部は、音声用マイクと雑音用マイクの入力信号を、共有メモリを介してノイズキャンセラコア部に伝達する。ノイズキャンセラコア部は、音声側入力に混入する雑音成分と、雑音側入力に混入する音声成分を、それぞれ個別の適応フィルタで推定し、推定結果を各信号から差し引くことにより雑音消去信号を生成する。雑音成分用と音声成分用の適応フィルタは、それぞれに専用のサブ適応フィルタを用いて推定した信号対雑音比でステップサイズを制御し、大きな消去量と小さな音声歪を両立する。雑音消去信号は、共有メモリ経由で雑音消去制御部に伝達される。ノイズキャンセラコア部と共有メモリ間の入出力信号転送は 512 サンプル毎に行う。総演算量は 78MIPS である。

5. おわりに

実環境で動作する音声対話機能を搭載した小型音声対話モジュールの機能、ハードウェア、及び実装形態について報告した。本モジュールは、現在、NEC を含む 5 機関のロボットに搭載され、機能検証中である。今後は、基本機能の安定性とモジュールの汎用性を向上させ、音声対話インターフェースの普及促進と知的対話の研究推進に貢献する予定である。

本研究の一部は、NEDO の次世代ロボット共通基盤開発プロジェクトの一環として行った。

参考文献

[1] M. Sato, A. Sugiyama, O. Hoshuyama, N. Yamashita, Y. Fujita, "Near-Field Sound-Source Localization Based on a Signed Binary Code", IEICE Trans. Fundamentals, Vol.E88-A, No.8, pp.2078-2086, Aug. 2005.
[2] M. Sato, A. Sugiyama, S. Ohnaka, "An Adaptive Noise Canceller with Low Signal-Distortion Based on Variable Stepsize Subfilters for Human-Robot Communication", IEICE Trans. Fundamentals, Vol.E88-A, No.8, pp. 2055-2061, Aug. 2005.
[3] Y. Fujita, "Personal Robot PaPeRo," J. of Robotics and Mechatronics, Vol.14, No.1, pp.60-63, Jan. 2002.
[4] 高野, 宇田, 石田, 西沢, 藤田, "ロボットプラットフォーム RoboStudio による RT コンポーネントの実現", Proc. RoboMec2007, 1P1-C02, May. 2007.
[5] http://www.w3.org/TR/speech-grammar/.
[6] 岩沢, "パーソナルロボット PaPeRo の音声認識インターフェース", 人工知能学会 AI チャレンジ研究会, Vol.13, pp. 17-23, Jun. 2001.