音声言語情報処理 12－12
（1996．7．27）

車載機器用音声対話システム

北岡教英　加藤利文　杉浦 恒　赤堀一郎
kitaoka@jo1.denso.co.jp
日本電装 (株)
〒448 愛知県刈谷市昭和町 1–1

カーナビゲーションシステムなど多くの機能を持った車載機器が増加している。これらのインターフェースとして、音声対話が有望である。

我々は、音声対話が車載用のインターフェースとなり得るためには、超大語彙認識が可能で、Uni-modal、Uni-mode な対話システムであることが条件であると考える。この考えをもとに車載機器用音声対話システムを開発した。システムは全国 15 万の地点を認識できる音声認識部と発声できる音声合成部、それらを制御する対話制御部からなる。

このシステムを被験者に使用してもらい、その様子を観察した。その結果、システムからの問いに対し、よくみられる応答のしかたや、慣れることによる応答の変化などが観測された。

A Speech Dialog System for In-vehicle Equipments

Norihide KITAOKA, Toshifumi KATO, Hisashi SUGIURA and Ichiro AKAHORI
NIPPONDENSO CO.,LTD.
1-1, Showa-cho, Kariya, Aichi, 448 Japan

In-vehicle equipments with complicated functions like navigation systems are on the increase. Speech dialog is a promising interface for these systems.

Speech dialog system in vehicle should be able to recognize very large vocabulary, should be Uni-modal and should be Uni-mode. We developed a speech dialog system in vehicle which satisfies these requirements. This system is composed of a speech recognition system, a speech synthesis system and a dialog control system. The recognition system can recognize 150,000 location names around Japan and the synthesis system can speak all these names.

We observed how subjects uses this system. We found some tendencies of responses to questions of system and subjects adjusts adjusted their responses as they got used to the operation.

1 はじめに

操作の複雑化している車載機器、特にカーナビゲーションシステムのインターフェースとして、音声対話入力が注目されてきている。我々は、カーナビゲーションシステムに音声対話を用いるうえで必要な条件を検討した。

カーナビゲーションシステムの重要な機能に地図表示とルート案内がある。これらの機能を音声で用いるには、日本中の地名の認識ができなければならない。したがって、条件の一つめとして、システムは地名の認識が可能となる、超大語彙の認識が必要となる。

二つめとして、注視を必要とする操作ができないことがあげられる。したがって、ボタンやタッチパネルなど他の方法と組み合わせることができない。

最後に、ユーザが何を言ったらよいかわからなくなり、途方にくれることがあってはならない。これは音声対話システムに一般的に言えることである。特に対話に意識を集中することができない運転中のユーザには重要な条件となる。

以下では、まず、上記を満たすために、システムの側に必要な条件を検討する。そして、その結果をもとに構築した音声対話ナビゲーションシステムを紹介する。さらに、それらを実際に使用してもらい、その時の様子を観察して得られた知見を述べる。

2 車載機器用の音声対話インターフェースに求められるもの

車載機器用の音声対話インターフェースには3つの条件として、

1. 超大語彙が認識できること
2. ボタンやタッチパネルなどと組み合わせてはならないこと
3. ユーザが途方にくれないこと

をあげた。本章では、その実現のために必要な要素について述べる。

2.1 大語彙認識

ナビゲーションシステムでよく利用される機能に、目的地の設定と現在地からのルート探索がある。しかし、その操作はボタン操作と画面の注視をともなうため、走行中は禁止される。

そこで、この操作が音声で行なえるようになることが期待される。この機能を満たすためには、十分詳細な地点の指定が容易にできなければならない。具体的には、市の下の町名のレベルや、町や村における大字のレベルまで入力できる必要がある。さらに、「京都府」「京都市」「左京区」のように、レベルごとに区切って発声することはユーザにとってわずらわしいので、ひと続きで発声できるのがよい。

われわれは、大語彙音声認識システム [1] を開発している。現在、このシステムは表1に示す地点にいくらかのコマンドなどを加えた、約15万語[1]が認識できる。地名については、郡名は省略可能であり、また「町」「村」の、「ちょう/まち」「そん/むら」の言い間違いは許している。施設名としては、役所、遊園地、ゴルフ場などが「県名 施設名」の形式で、インターチェンジなどが「道路名 インターチェンジ名」の形式で入力できる。

表 1: 認識可能地名

<table>
<tr><td rowspan="5">都道府県
47
(108166)</td><td rowspan="2">市
665
(66956)</td><td>区
149 (11750)</td><td>町
11601</td></tr>
<tr><td>町
54541</td><td></td></tr>
<tr><td>郡
563 (40159)</td><td>町村
2578(39596)</td><td>大字
37018</td></tr>
<tr><td>東京23区
23 (966)</td><td>町
943</td><td></td></tr>
<tr><td>島嶼
1 (38)</td><td>町村
9 (37)</td><td>大字
28</td></tr>
<tr><td colspan="4">施設名 (インターチェンジ、病院など)
36419</td></tr>
</table>

数字は個数. () 内の個数には下位階層の地名が含まれる.

このシステムの認識率評価実験をワークステ

[1]ここでは、「愛知県」も「愛知県刈谷市昭和町」も単語であるとする。

ーション上で行った。認識対象語彙は、市の下のレベルおよび大字レベルまでの地名(途中のレベルまでの発声も可)108,166 である。また、実験サンプルは、1 レベル (都道府県), 2 レベル (市郡まで), 3 レベル (町村まで) 各 50 地名、計 150 地名を男性 11 名、女性 10 名が発声した、男性 1650 サンプル、女性 1500 サンプルである。結果を表 2に示す。

表 2: 地名認識システムの性能 (認識率)

	男声	女声
1 位認識率	94.4%	92.7%
3 位以内認識率	98.5%	97.5%

2.2 Uni-modal, Uni-mode

現在主流の入出力装置として、キーボード、ボタン、タッチパネル、ディスプレイがある。音声認識/合成は、これらと組み合わせたマルチモーダル (Multimodal) インターフェースとすることで、効率的で便利になると一般的には考えられている。地名入力に関しても、タッチパネルやキーボード、マウスと組み合わせて入力を効率化できることが報告されている [1] [2]。しかし、車載機器の操作には、目を使用する他の方法をとれないため、音声認識/合成のみの **Uni-modal** インターフェースとなる。

音声入力には誤認識が不可避であり、それを考慮に入れた場合、確認などの過程が必要である。車載においては、それは音声対話で行なうのが妥当である。

一般に対話的な入力システムは、対話の進行につれて入力モードを切替えてユーザの入力を想定する。しかし、入力方法が音声の場合、想定から外れた入力がなされることが非常に多い。

例えば、地名を設定する場合として、次の対話例 1 をあげる。

対話例 1

ユーザ: 「愛知県刈谷市昭和 (ショーワ) 町」
システム: 「愛知県刈谷市松栄 (ショーエー) 町ですか？」
ユーザ: 「(*)」

システムは (*) として、「はい」または「いいえ」を想定するモードに移ったとする。ボタンや画面とのマルチモーダルシステムならば、画面にこれらの回答を表示すると同時に回答のボタンを用意する。ユーザはそれらから選択するので、想定からはずれた回答がなされることはない。

ところが、音声対話の場合には、「愛知県刈谷市昭和町」と繰り返すケースや、「ちがう」など、同じ意味で異なる発声をするケースなど、想定外の発声も非常に多く見受けられる。

この原因として、ユーザから見たシステム (メンタルモデル) が、実際のシステムからずれてしまうことがあげれる。これは、対話戦略をうまくたて、完全にユーザを誘導することで解消される。しかし、十分な情報を含ませるためには、システムは長い発声が必要となる。例えば、対話例 1 のシステムの発話も、「愛知県刈谷市松栄町ですか?『はい』または『いいえ』で答えて下さい。」のようになる。このような長いシステムの発話は、ユーザにわずらわしさを感じさせる。また、運転中にシステムの発声に聞き入ったり、モードの遷移を意識することは好ましくない。

また、想定外の発声をされた場合、現在のところ「想定外であること」を認識する手法が確立していない。その結果、システムは「とんちんかんな」返答をすることになり、ユーザはどうしてよいかわからなくなって途方にくれてしまう。

したがって、システムは、ユーザから見てモード切替のない、つまり **Uni-mode** なシステムであることが望ましい。

3 音声対話システムの構築

ここまでの考察をもとに、実際に音声対話システムを構築した。

このシステムでは、音声により、地図の拡大・縮小などのナビゲーションシステムの操作が行える。特に、目的地の設定において、入力地名の確認、認識誤りの修正および確定が、対話に

よって行われる。

対話処理部の内部状態オートマトンは、主にメインの状態と地名確認状態からなる。しかし、Uni-mode システムの方針から、認識可能な語彙は常に同じである。

特に、認識誤りの場合には、ユーザは「いいえ」と答えることで、直接認識誤りであることを示すことができる。この場合、システムは地名認識時の第 2 候補を示し、確認を求める。また、ユーザが地名を言い直した場合には、認識除外単語リスト (後述) を用いた認識・対話処理により対話処理部が再度ユーザに確認を求める。

3.1 音声対話システム全体の構成

全体図を図 1に示す。

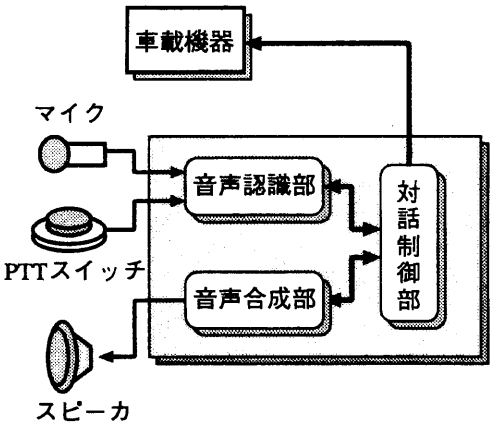

図 1: システムの全体構成

音声認識部は、対話制御部からの指示により認識処理を行い、結果を対話制御部に返す。対話制御部は、認識結果および自身が管理する内部状態から、音声合成部への発声の指示や車載機器の操作などのアクションを実行し、自身の内部状態を遷移させる。同時に、認識部に認識開始の指示をする。

音声認識部として、2.1節で説明した音声認識システムを用いる。ユーザは、PTT(Push-to-Talk) スイッチを押しながら音声を入力する。

音声合成部は、全国の地名を発声する必要があるため、テキスト合成方式を採用した。

3.2 対話制御部の特徴

Uni-mode システム

対話処理部は内部状態の表現としてオートマトンを用いる。一般にオートマトンを用いた対話システムは、オートマトンの 1 状態がシステムのモードに対応する。しかし、ユーザは状態の遷移すなわちモードの遷移を常に意識していなければならず、それを見失うと途方にくれてしまう。

このようなことに陥らないため、われわれは、図 2 のようなオートマトン作成方針をとった。内部的には複数の状態 (Multi-state) を表現するが、ユーザから見た場合には単一のモード (Uni-mode) に見えるようにする。つまり、状態によって認識語彙は変わらないが、語彙とアクションの対応が変わる。

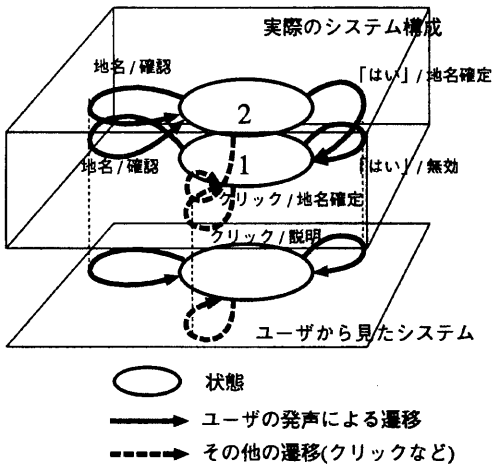

図 2: Uni-mode システムの概念

認識除外単語リスト

われわれはこれまでに、地名認識のような大語彙の認識システムの場合、画面に上位候補のメニューを表示してタッチパネルなどで選択させる方式 (メニュー選択方式) を併用することにより十分な認識率を得ることができ、有効であることを示した [1]。しかし、車載機器の場合には画面を用いることができず、この方式は採用できない。

システムがユーザに認識結果の確認をもとめ、ユーザが「いいえ」と発声するなど、認識誤りであることを示した場合に、順次候補を提示することで、メニュー形式と同等の認識率を得ることができる。しかし、「いいえ」を連続で発声し順次候補を確認する作業は、特に慣れたユーザには、まわりくどさを感じさせる。

また、対話例1であげたように、ユーザはすぐに言い直しをしてしまう場合が多い。そこで、確認を求める場面でも地名の言い直しを許すことにする。しかし、特定の単語に誤る傾向が強い語は、入力し直してもやはり同じ誤りを繰り返し、なかなか入力できない場合が生じる。

そこで、認識除外単語リストを導入する。例えば対話例1における(*)で、ユーザが言い直しをした場合を考える。対話制御側は、先のユーザー発話の認識結果となった「愛知県刈谷市松栄町」を認識除外単語リストに登録しておく。認識システムは同じ認識誤りを起こし、再度1位候補として「愛知県刈谷市松栄町」を返すことがある。対話制御部は、このリスト中にある結果を除外する。この結果、「愛知県刈谷市松栄町」は除外され、2位認識候補が1位に浮上する(図3)。

対話制御部は、地名が確定した場合や、他のコマンドが実行された場合などに認識除外単語リストをクリアする。

この手法により、複数候補選択方式と同等の認識性能を得ることができる。

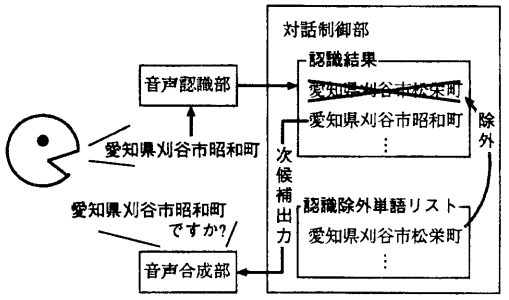

図3: 認識除外単語リストの働き

発声以外による状態遷移

システム内の状態遷移は、原則としてユーザの発声により起こるが、本システムでは、以下の条件によっても状態遷移が起こる。

- 時間経過
  ある状態に一定時間とどまった後、自動的に遷移する。これは、ユーザが対話を途中でやめても自動的に初期状態にもどる場合などに用いる。
- 音声入力スイッチのクリック
  頻繁に用いる「はい」などの発声は、なれたユーザにはわずらわしくなる。そこで、音声以外の入力手段として、音声入力スイッチをクリックできる。

音声入力開始による音声出力の中断

音声出力中にPTTを押すことにより、音声出力を中断する。これにより、音声出力の終了を待つことなく入力が開始できる。

4 対話事例

音声対話ナビゲーションシステムを室内に設置し、システムの使用経験のない被験者(男性17名、女性2名、全員20才台)にユーザとして操作してもらった。

ユーザには、コマンドのリストと地名や施設名の入力方法を書いた説明書を渡した。そして、実験者は1分程度の簡単な説明をした後、1度だけ地名の入力を実演した。確定には「はい」を用いたが、クリックできることも伝える。誤認識の訂正は、「いいえ」または、言い直しができることを伝えたが、実演していない。その後すぐ、ユーザに自由に操作してもらった。

このとき現れた特徴的な対話例をあげる。ただし、状況に応じて、実験者はユーザに対して説明を行っている。

対話例2

ユーザ: 「名神高速道路 西宮北インターチェンジ」
システム: 「名神高速道路 一宮インターチェンジですか?」
ユーザ: 「西宮北インターチェンジ」

前半部分が正解で、後半のみ誤認識されたとき、ユーザはとっさに後半部分のみを訂正してしまった。

対話例 3

ユーザ: 「愛知県岡崎市細川町」
システム:「愛知県岡崎市細川町ですか？」
ユーザ: 「はい、そうです。」
実験者: そうです、は、やめてください。
ユーザ: 「愛知県岡崎市細川町」
システム:「愛知県岡崎市細川町ですか？」
ユーザ: 「はい、そうです。」

システムが対応できない返答に対し、実験者がその返答を使用しないよう要求した。しかし、直後の対話において、ユーザは全く同じ返答をしてしまった。このユーザは、このように答えられないと非常に苦痛を感じるそうである。

対話例 4

ユーザ: 「愛知県岡崎市美合 (ミアイ) 町」
システム:「大分県大分市都町ですか？」
ユーザ: 「**愛知県岡崎市美合町**」
システム:「愛知県岡崎市宮石 (ミヤイシ) 町ですか？」
ユーザ: 「いいえ。」
システム:「愛知県岡崎市宮地 (ミヤジ) 町ですか？」
ユーザ: 「いいえ。」
システム:「愛知県岡崎市美合町ですか？」
ユーザ: 「はい。」
システム:「目的地を設定しました。」

一連の対話中で、誤認識したシステムの確認に対して、地名を言い直す場合と、「いいえ」と答える場合がみられる。誤認識の度合が大きい (この場合、県から異なっている) 場合に言い直しを、小さい (最後の階層のみ異なっている) 場合には「いいえ」を用いている。誤認識の度合で返答が異なる傾向は、このユーザに限らず一般的にみうけられた。

その他の気付いた点

- ユーザは操作になれるにしたがって、
  - ·「いいえ」よりも言い直しを多用する。
  - ·「はい」よりもクリックを多用する。
- 「えー」などの不用語が地名の先頭や地名階層の境界で現れることがあった。
- 認識対象よりさらに下層の地名 (字名や番地など) まで発声する場合があった。
- 施設名はすべてを認識対象としていない。そのため、ユーザは目的の施設が入力できるかどうかわからなくなり、不安になる。
- 確認の「はい」の誤認識は、タスク達成にとって致命的である。正解が認識除外単語リストに登録されるため、何度言い直しても認識されなくなり、ユーザは途方にくれる。

5 おわりに

車載機器用の音声対話システムの条件を検討し、音声対話ナビゲーションシステムを構築した。そして、ユーザがシステムを操作する様子を観察した。

今後の課題として、ユーザの傾向を反映して、より使いやすいシステムとすることがあげられる。具体的には、語彙の充実や、階層途中からの入力 (対話例 2 の最後の発話など) を可能にするなどがある。不用語や未知語への対策も必要である。さらに、走行中の操作を観察する必要がある。

参考文献

[1] 赤堀一郎, 加藤利文, 北岡教英. "地名認識システムとその応用", 情処研報, 95-SLP-7-9, pp.55-60, 1995.

[2] 荒井和博, 吉岡理, 管村昇, 嵯峨山茂樹. "マルチモーダルインタフェースを持つ住所入力システムの評価実験", 情処研報, 95-SLP-7-19, pp.107-112, 1995.

[3] 嵯峨山茂樹. "なぜ音声認識は使われないか・どうすれば使われるか?", 情処研報, 94-SLP-1-4, pp.23-30, 1994.