

カーナビゲーション向け音声認識の 新たな展開

成田知宏* 古本友紀***
岡登洋平** 井上 譲***
花沢利行*

New Developments in Speech Recognition for Car Navigation System

Tomohiro Narita, Youhei Okato, Toshiyuki Hanazawa, Yuki Furumoto, Yuzuru Inoue

要 旨

三菱電機では安全・安心な操作を実現するために、2001年からカーナビゲーションに音声認識機能を搭載している。2004年に発売した“CU-H9000”では、施設名のあいまい認識を導入することで省略、語順の入れ替え、同義語を含む言い換え語による施設名の認識を実現した⁽¹⁾。しかし従来の音声認識方式では、認識語彙(ごい)数が増加するにしたがって演算量が大幅に増加するという問題があり、次の点でユーザーの利便性が低下していた。

- (1) 施設件数を10万件程度に制限するため、音声認識対象が駅や遊園地等のランドマーク的な施設に限定される。
- (2) 操作コマンドはマニュアル記載の表現(“近くのコンビニ”等)しか登録されていないため、異なる表現の発話は認識できない。

そこで当社では、どの音に近いかという音響的な探索と、どの言葉に近いかという言語的な探索を分離するアプローチを採用した。単純な文字列変換処理としての音声認識部

と、音声認識のポスト処理としての検索部に分離することで、語彙数が大幅に増加した場合でも演算量の増加を抑制することができる。具体的には(1)の課題に対して、統計言語モデルを用いた音声認識と高速テキスト検索技術を適用し、100万件規模の超大語彙の音声検索を実現した。(2)の課題に対しては、ユーザーが用いる操作コマンドの多様な表現を収集し、統計言語モデル及びキーワードとコマンドの対応関係を学習することで、認識結果に含まれるキーワードからコマンドにマッピングする手法を開発した。

このように音声認識とポスト処理を組み合わせることで、従来にない機能を実現することは、カーナビゲーション向け音声認識の新たな展開を示している。今後は、意味理解処理や対話処理を組み合わせた音声ヘルプ機能や、ユーザーの嗜好(しこう)性分析処理を組み合わせた知的な情報検索・提示機能の実現を進めていく。



HDDカーナビゲーション “NR-HZ001VDT”

全国の主要施設約1,000万件を簡単に検索できる新開発の“超大語彙音声認識技術”を搭載したほか、うっかり不要な言い回しをつけて発話しても音声コマンドを抽出できる“キーワード抽出”を搭載している。更に、音声認識・タッチパネル・リモコンが連動して動作するマルチリレーオペレーションを搭載することで操作性を大きく向上させている。

1. ま え が き

当社では、カーナビゲーション向け音声認識機能のユーザー満足度向上のため、発話タイミングや語彙外発話などの問題点の改善を継続的に行っている。現状の大きな課題の一つとして、ユーザーが認識対象外の語彙を発話してしまう語彙外発話がある。この問題に対しては、あいまい認識⁽¹⁾を搭載することで対応しているが、認識対象ジャンル(デパート、レストラン等の分類)を拡大すると認識語彙数の増加に伴い、言い換え語の追加・整備コスト及び演算量が増加するという問題がある。また操作コマンドについては、マニュアル記載の表現しか登録されていないため、異なる表現の発話は認識できないという問題がある。

これらの問題に対して当社では、どの音に近いかという音響的な探索と、どの言葉に近いかという言語的な探索を分離するアプローチを採用した。単純な文字列変換処理としての音声認識部と、音声認識のポスト処理としての検索部に分離することで、語彙数が大幅に増加した場合でも演算量の増加を抑制することが可能になった。

本稿では開発した音声認識技術とポスト処理技術、及び製品に搭載した音声インタフェースについて述べる。

2. 音声認識とポスト処理技術

2.1 統計言語モデルを用いた音声認識

言語モデルとは、認識対象語彙の情報を保持するデータである。音声認識は言語モデルに登録した語彙の中で、入力音声に対するもっともらしさ(認識スコア)が一番高いものを認識結果として出力する。このため、施設名として正式名称の“東京国立近代美術館”だけを言語モデルに登録すると、“東京近代美術館”のように“国立”を省略した発話は認識できない。対策として先に述べたあいまい認識では、施設名を“東京”“国立”“近代”“美術館”というように単語に分割し、各単語とともに単語間の接続規則を言語モデルに登録している。例えば“国立”は省略可能という規則を登録すれば、“東京近代美術館”という正式名称以外の発話も認識可能となる。しかし、多様な呼称をカバーする規則を整備するのは困難でコストが大きいという問題があった。

これに対して今回開発した音声認識方式では、単語間の接続のしやすさを言語スコアと呼ぶ数値で表す統計言語モデルを用いた。例えば施設名の場合、“東京”という単語の次に“国立”や“近代”が続く言語スコアがそれぞれ7、2等と登録されており、各認識候補の認識スコアには、この言語スコアが加算される。言語スコアは大量の施設名から自動的に算出が可能のため、小コストで言語モデルを作成できる。また正式名称に出現しない単語接続に対しても、近似的な言語スコアを与える方法を開発しており、正式名称以外にも認識可能である。

ただし施設名が数十万語の規模になると、施設名に含まれる単語数も数万語のオーダーとなるため、従来のあいまい認識と同様に音声認識時の演算量とメモリ量が問題となる。そこで、各単語をより小さな単語に分割して表現する。例えば“美術館”は、“びじゅつ”と“かん”という単語で表現する。単語の総数は施設名中での出現頻度の高い上位5,000個程度に制限する。単語の連鎖で表現できない場合は、音節の連鎖で表現する。例えば“近代”は、“き”“ん”“だ”“い”という音節の連鎖で表現する。このように音節を言語モデルの単位とすることで、検索対象語彙数が増加しても認識単語数を一定数に保ち、リソースに制限のあるカーナビゲーション上での大語彙認識を可能にしている。

2.2 高速テキスト検索手法による施設名検索

開発した施設名検索は図1に示すとおり、音声認識部とテキスト照合部で構成される。音声認識部では2.1節で述べたとおり、検索対象語彙を単語と音節に分割して認識する。認識結果は登録語彙と完全一致しない場合が生じるので、テキスト照合部で認識結果の音節列と最も良くマッチする施設名を検索する。具体的には図2に示す転置インデックスを用いて、各施設名に対する検索スコアを算出し、スコア最大の施設名を検索結果とする。索引語としては図2に示すとおり2音節の連鎖を用い、索引語ごとに、その索引語を含む施設名のID番号を保持する。

例えば、認識結果の音節列が“しいばら”の場合、“しい”

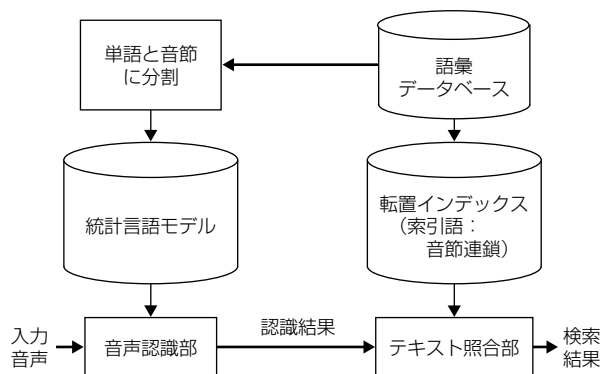


図1. 施設名検索方式

語彙データベース	
語彙ID	語彙の音節列
0	よこはますらしあ
1	よこはまぎんこう
2	しいばらだいず

↓ 自動生成

転置インデックス	
索引語: 音節連鎖	語彙ID
よこ	0,1
こは	0,1
はま	0,1
しい	2
いば	2
ばら	2
...	...

図2. 転置インデックス

“いば”“ばら”の3種類の音節連鎖が存在する。次に転置インデックスを参照して、これらの音節連鎖を含む施設名に対し、音節連鎖1個ごとにスコアを1点加算する。図2の例では、施設ID=2“しいばらだいす”のスコアが3となり、検索結果の第一位として出力される。このように転置インデックスを参照することで、高速にスコアを計算することができる。なお音節連鎖を索引語とする利点は、単語や単語連鎖と比較して音声認識誤りに頑健なことである。例えば“東京近代美術館(とーきょーきんだいびじゅつかん)”という発話に対し、音声認識部の出力結果が“東京近代美術館(とーきょーきんだいぶじゅつかん)”となった場合、単語単位では誤認識した“美術館”が“美術館”とはマッチせずスコアに寄与しないが、音節単位なら“じゅつ”“つか”“かん”の3個の音節連鎖が“びじゅつかん”とマッチするため、正解の“東京近代美術館”のスコアに寄与できる。

2.3 キーワード抽出型コマンド

施設名検索と同様の手法を適用し、音声コマンドに対するユーザー発話の受容性を高めたキーワード抽出型コマンドを実現している。

キーワード抽出型コマンドの処理を図3に示す。音声認識部は、入力音声を変換して単語列を出力する。テキスト照合部は、認識結果に含まれるキーワードに基づいて文例データベースを読み出すとともに、文例ごとのスコアを計算する。それぞれの文例には、該当するコマンドが対応付けられており、スコア上位のコマンドを出力する。フィルタ部は、現在のナビゲーション状態に応じて有効なコマンドを抽出し、結果を確定する。図3の例では、入力発話“コンビニ近くにある”を音声認識し、テキスト照合部はキーワードの長さ、入力発話の長さ、文例の長さを考慮したスコアを算出する。この例では2件のコマンド候補を抽出している。フィルタ部は、ナビゲーションの状態を考慮して最終的なコマンド“周辺検索(コンビニ)”を決定する。この方式ではキーワードを単位とした照合を行う。このため、表現の追加(例えば“～して下さい”などの丁寧表現)、助詞の欠落、語順の入替えが生じてコマンドを推定可能

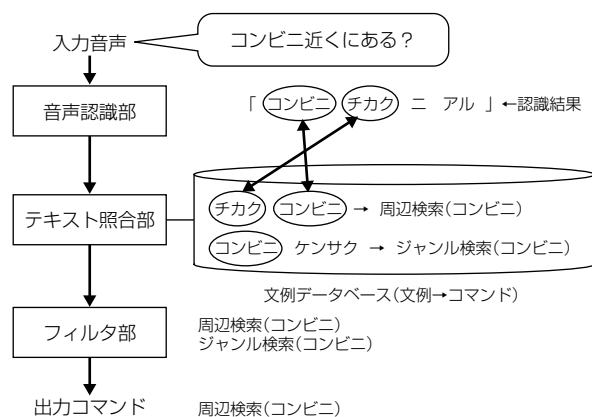


図3. キーワード抽出型コマンドの処理

であり、従来方式と比べて受容性が拡大する。

一方、コマンド表現で学習した統計言語モデルを用いると、待ち受け対象外の発話をコマンドに誤る可能性が高くなる。予備調査の結果、カーナビゲーション地図表示画面における待ち受け外発話の8割は施設名であった。そこで、リジェクト用として、施設名に類出するキーワードと音節列を統計言語モデルの学習データに追加した。さらに、施設名キーワードの検出時に、施設名検出を通知する特別なコマンドを出力するようにした。これによって、後述する施設名リジェクト機能を実現している。

3. 市販カーナビゲーション“NR-HZ001シリーズ”への適用

2章で述べた音声認識とポスト処理を適用した、HDD(Hard Disk Drive)カーナビゲーションの音声インタフェースについて述べる。

3.1 音声50音検索

音声50音検索とは、検索したい施設名の読みを手操作で一文字ずつ入力する代わりに、施設名を声で入力し検索を行う機能である。50音パレット(図4)で入力する場合と比較して、極めて高速な入力が可能となり、ユーザーの利便性を大幅に向上させる機能である。前述の統計言語モデルを用いた音声認識処理と、高速テキスト検索技術の採用によって、手入力による検索が対象とする全国1,000万件の施設に対する音声施設検索を可能としている(同時に検索対象とする施設は同一都道府県内に限定し、最大100万件である)。

この機能の動作を“梅田スカイビル空中庭園展望台”が検索対象である場合を例に説明する。

50音検索画面(図4)で、“声で入力”をタッチすることで音声認識処理が起動する。そして、ユーザー発話“スカイビル”を認識した時の画面が図5である。ユーザーがシステムの動作を把握できるよう、音声認識処理の結果“1>スカイビル”が表示されている(同時に音声出力も行われている)。またテキスト検索処理の結果として、自車位置の都道府県内にある施設の中で“スカイビル”を名称の一部に



図4. 50音検索画面(50音パレット)



図5. 検索結果リスト画面

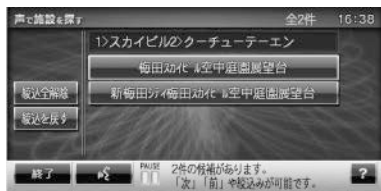


図 6. 検索結果リスト画面(絞り込み時)



図 7. 都道府県切替え画面

含むものが48件検索され、その一部がリスト表示(“きじスカイビル店”“スカイビル駐車場”など)されている。

このリストが表示されたところで、ユーザーは手操作又は“次”“前”の発話によって、施設を選択することができる。また、施設名を音声入力で追加することで候補を絞り込むこともできる。図 6 は“クチュウターエン”を追加で発話して、検索候補の絞り込みを行った場合の画面である。検索対象を別の都道府県の施設に切り替える場合は、図 7 の都道府県切替え画面で、手操作又は音声操作で希望する都道府県を選択することができる。

3.1.1 検索対象施設の拡大と都道府県別検索

従来の音声施設検索は、手操作による検索より大幅に少ない施設しか検索対象としていなかった。そのため、手操作で検索できる施設が音声では検索できない場合があり、音声施設検索に対するユーザーの信頼度を低下させる原因の一つとなっていた。

音声50音検索では、手操作による50音検索と同じ全国1,000万件の施設を検索対象とすることで、この問題を回避している。しかし全国の施設を検索対象とした場合は、同一又は類似した名称を持つ施設が大量に検索されるため、逆にユーザーが希望する施設を見つけ難くなるという別の問題が生じる。そこで検索対象を自車位置の都道府県に限定し、別途対象都道府県切替えの操作画面を用意することで対応している。検索対象とする施設数を制限することで、音声認識精度と検索速度の向上にも貢献している。

3.1.2 音声認識結果の表示

音声認識の問題点の1つに、ユーザーが同じように発話しても、声の調子や背景騒音の影響で常に同じ認識結果が得られるわけではないという不安定さがある。そのため音声認識結果がそのまま検索結果となっていた従来の音声検索では、音声検索のたびに表示される施設が異なることがあった。

今回開発した音声50音検索では、2.2節で示した認識誤りに頑健なテキスト検索技術を用いているため、この問題に起因する使いにくさが低減されているが、極端に間違っ

た音声認識結果からは正しい検索結果を得ることはできない。また音声認識結果が正しくても、ユーザーが希望する施設が検索対象に含まれていない場合もある。そこで自分の発話がどのように音声認識されたか、どのような認識結果(読み)でテキスト検索されたかをユーザーが把握できるように、ユーザー発話に対する認識結果の表示・ガイダンスを行っている(図 5, 図 6)。これによってユーザーは、音声検索の結果に対するその後の対応(再度音声検索をやり直すのか、絞り込みを行うのかなど)を適切に決定できる。

3.2 施設名リジェクト

カーナビゲーションにおける音声認識では、表示画面やナビゲーションの内部状態、ユーザーの操作などに応じて、認識対象となる語彙が設定されている。しかし、ユーザーはどのような語彙が認識されるのか分からないため、しばしば認識対象外の語彙を発話してしまう。このような発話は正しく認識されず、ユーザーの期待するナビゲーション動作も行われなため、音声操作に対するユーザー不満の原因の一つとなっていた。

この問題への対応策として、市販カーナビゲーション(2009年秋以降のモデル)には、キーワード抽出型コマンドを応用した施設名リジェクトの機能を導入している。これは地図が表示されている場合や、施設名の音声認識が実行できない状況で、ユーザーが認識対象外の施設名を発話したことを検知すると、施設名発話が現在認識対象ではなく、50音検索画面で認識可能であることを伝える機能である。この機能によって、施設を検索したいユーザーを音声50音検索画面に誘導することができる。

4. む す び

音声認識とポスト処理を組み合わせることで新機能を実現する、カーナビゲーション向け音声認識の新たな展開を示した。音声は人間にとって最も自然な入力インタフェースであるが、音声認識だけでは単なる入力手段に過ぎない。カーナビゲーションの操作性向上のためには、本稿で述べたように、ポスト処理を組み合わせた利便性の高い機能を実現することが必要である。今後は、意味理解処理や対話処理を組み合わせた音声ヘルプ機能や、ユーザーの嗜好性分析処理を組み合わせた知的な情報検索・提示機能の実現を進めていく所存である。

参 考 文 献

- (1) 岩崎知弘, ほか: カーナビゲーションシステムの音声インタフェースー現状と将来展望ー, 三菱電機技報, 78, No.9, 613~617 (2004)
- (2) 花沢利行, ほか: 統計言語モデルによる音声認識とテキスト照合に基づく大語彙音声検索, 日本音響学会研究発表会講演論文集, 61~62 (2009)