音声言語情報処理 7－9
（1995．7．21）

地名認識システムとその応用

赤堀一郎　加藤利文　北岡教英
akahori@jo1.denso.co.jp　ribun@jo1.denso.co.jp　kitaoka@jo1.denso.co.jp
日本電装 (株) 情報システム技術部
〒448 愛知県刈谷市昭和町 1–1

我々は, 音声認識が使われるためには, 認識できる文・単語の範囲が明確であること, キーボードなど他の入力手段より音声の方が簡単に入力できることなどが重要であると考えている. この条件にあてはまる例として, 日本全国の約 10,000 個所の地名が認識できる音声認識システムを開発した. 連続出力分布型 HMM を用いた不特定話者・連続単語認識によって, 男女の平均で, 1 位認識率 94%, 5 位以内認識率 99%を得ている. DSP を 2 個使用したハードウェアにより, リアルタイムの認識が可能である. 本稿では, このシステムの認識アルゴリズムとハードウェアについて報告する. また, 公衆情報端末 (据置き型の情報端末) にこのシステムを組み込んで評価実験を行った結果を示す.

Place Name Recognition System and its Application

Ichiro AKAHORI, Toshifumi KATO and Norihide KITAOKA
Information & Communication System Engineering Dept.,
NIPPONDENSO CO.,LTD.
1–1, Showa-cho, Kariya, Aichi, 448 Japan

It is important for a speech recognition system to be used that the sentences/words which the system can recognize are clear and that the input operation with voice is easier than with other methods. We have developed a speech recognition system which can recognize 10,000 Japanese place names, which satisfies these conditions. This can deal with voices of both sexes and achieves the recognition rate of 94% with the first candidate and of 99% with the first 5 candidates. This can make real-time recognition on a hardware with two DSP's. We report on the algorithm and the hardware. We evaluated the system applied to public information terminal as a place-name input device. We also report the result.

1 はじめに

電子機器が高度化するにつれ, その操作方法が複雑化する傾向にある. 音声入力は操作を容易にするための有望な手段である. 音声入力は入力速度が速い, 習得が容易といった利点の他に, 手や目を使う他の作業を妨害しないという利点がある. これは, 自動車のおけるユーザインタフェースとして, 安全性の向上に継がる重要な性質であり, 音声入力の実用化が望まれている.

しかし音声認識の実用化に向けての様々な努力が続けられているが, 現状では音声認識はごく限られた分野でしか応用されていない. 音声認識の応用が広まらない理由として, 認識率が低い, ユーザインタフェースが未熟, 応用についての考察が不足などの点があるといわれる [1].

我々は, 音声認識技術と応用分野の組み合わせをうまく選択することで, 役に立つ音声認識応用システムを実際に構成できると考えている. その実証として, 地名認識システムを開発した. これは, HMM(Hidden Markov Model) を用いた不特定話者・連続単語認識によって, 日本全国の約 10,000 個所の地名を認識するシステムである. このシステムは, カーナビゲーションの経路案内における地名設定などに利用可能である.

本稿では, はじめに「どうしたら役に立つ音声認識応用システムが作れるか」について考察し, 次に開発した地名認識システムのハードウェア, アルゴリズムについて紹介する. 最後に, このシステムを使用した音声入力公衆情報端末 (据置き型情報端末) と, その評価実験の結果を示す.

2 役に立つ音声認識応用システムを作るには

音声認識があまり応用されないのは, 音声による入力よりキーボードやボタンなど音声以外の入力手段の方が簡単で快適だからである.

音声認識を役立たせるためには, 「音声でもできる」ではなく「音声を使った方が簡単だ, 快適だ」という視点から応用システムの開発を進める必要がある.

2.1 タスクの選定

どのようなタスクに対しても音声認識が有効ということはありえない. 認識率や認識語彙数などの音声認識性能が上がるほど, 有効性のあるタスクの範囲が広がるが, 現在の不十分な認識性能では適用するタスクを注意深く選定することが重要である. このとき, 次のことについて注意する必要があると考える.

メンタルモデルのずれ

音声認識システムにとって, どのような文・単語でも認識できることは理想であるが, これは実現困難であり, 現実的には認識対象となる文・単語の範囲を限定する必要がある. 一般に, これはタスクを限定することで行われるが, 情報検索などのようにある程度「知的」なタスクになると, 実際の限定範囲とユーザが感じる限定範囲 (メンタルモデル) とのずれが大きくなってくる. このため, ユーザが認識できるはずだと考えた文・単語が認識されないことがしばしば生じ, これが使い難さの原因となっていると考えられる.

語彙数

一方, 語彙数の少ない (10 語程度) タスクでは, このようなずれはほとんど発生しないが, この場合, 音声よりキーボードやボタン類などの他の方法を使った方が簡単に入力でき, 音声認識の高速性を活かせない. 語彙数が多く (100 語〜) なってくると, キーボードやボタンでは選択に必要となる操作回数が多くなり, 相対的に音声認識の優位性が増してくる.

したがって, 音声認識が有効であるためには, タスクの規模が大きく (認識語彙数が大きく), かつユーザが考える認識可能語彙と実際のそれとが一致する (「知的」でない) タスクを選択しなければならない.

限定範囲のシステム側とユーザ側のずれが少なく, 規模が大きいタスクの例として, 我々は地名の認識を考えた. 地名を対象としているので, 何が認識でき, 何が認識できないかをユーザが容易に把握することができる. また日本全国の約 10,000

個所の地名が認識対象であり比較的大規模なタスクであるので, 他の入力手段に比べて音声認識が優位になると考えられる. このシステムについて3節で述べる.

2.2 ユーザインタフェースの工夫

音声認識において誤認識は避けられない. ユーザインタフェースの工夫で誤認識による不快感を低減することも重要である. 毎回正否の確認を行う方法は煩わしい. 人間同士のように「知的」な対話によって対処する方法が望まれるが, 先の述べた「知的」なタスクにおける限定範囲のずれの問題が発生する. また, これに必要となる意味理解の技術もまだ確立していない. 音声認識のみにこだわらず, 快適性が増加するならばキーボードやボタンなど他の手段の積極的な併用を考えることも必要であろう.

1 位認識率 (認識順位が 1 位になる確率) があまり高くなくても, n 位以内認識率 (認識順位が n 位以内になる確率) は十分高く, 誤認識が実用上無視できるレベルであるとする. このとき, 1 位から n 位までの候補をすべて同時にユーザに示し, キーボードなどによって選択させる方法が考えられる. この方式では, 正解が n 位までに入れば正しく認識されたと考えることができる.

つまり音声認識を非常の多くの選択肢を少数に絞り込む手段として捉えることにより誤認識の定義が変わり, 見掛け上認識率を上げることができる. このようにしても, 本質的な認識性能には変化はないが, ユーザに与える印象がよくなると思われる.

4節で紹介する音声入力公衆情報端末では, 第 5 位までの候補を (認識順位順ではなく) 辞書式順に並べて表示し, それをタッチパネルで選択する方式を用いている.

3 地名認識システム

我々が開発した地名認識システムの認識アルゴリズム, ハードウェアについて説明する. このシステムは, HMM を用いた不特定話者の連続単語認識によって, 日本全国の約 10,000 個所の地名を認識するものである.

3.1 認識可能地名

認識できる地名は表 1 に示すように, 日本全国の約 10,000 個所の地名である. 表の斜線部の地名は愛知県内の地名に限られる. 地名の発声は都道府県名から行う必要がある (例: 「北海道札幌市」○, 「札幌市」×) が, 愛知県内の地名については県名が省略可能 (例: 「刈谷市昭和町」) である.

表 1 認識可能地名

都道府県 47 (10111)	市 662 (5798)	区 125 (1586)	町 1461
		町 3550	
	郡 570 (4233)	町村 2577 (3663)	大字 1086
	東京23区 23		
	島嶼 1 (10)	町村 9	

数字は個数. () 内の個数には下位階層の地名が含まれる. 斜線部は愛知県のみ

3.2 システム構成

このシステムのハードウェアは, 図 1 に示すように TI 社製の DSP(TMS320C31, 50Mflops) を二個使用したものである. 基板のサイズは 250mm × 170mm である.

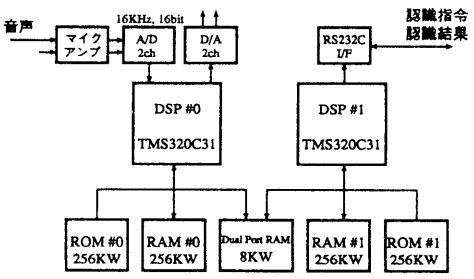


図 1 ハードウェア

DSP#0 で音声の特徴抽出と出力確率の計算を行い, DSP#1 で One Path DP による認識処理を

行っており, 発話終了後, 1～4 秒で認識結果が得られる. 外部との通信は, RS232C で行っている.

図 2 に示す処理をこのハードウェア上で実行し, 地名認識を行う.

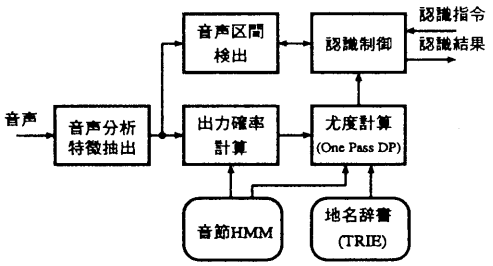


図 2 システム構成

3.3 音節 HMM

このシステムでは音節を単位とした認識を行っている. 音節は, 日本語の 110 音節 (外来音含む) に促音と無音を付加し, 計 112 個とした.

各音節は状態数 4 の連続分布 HMM で表現した (ただし, 促音と無音の状態数は 1). left-to-right 型であり, 次々状態への遷移は持たない. 状態遷移の遷移元で出力確率分布は結ばれており, 分布の個数は 4 である. 出力確率分布は 32 次元対角ガウス分布としている. 混合数は 2 であり, 各々の分布は, 男声と女声に対応している.

3.4 音節 HMM の学習

音節 HMM の学習は連結学習 [2] によって行った. 学習用音声データとして, 男性音声には電子協日本語共通音声データの地名 (100 地名×56 名) と日本音響学会研究用連続音声データベースの ATR 音素バランス文 4518 発声 (503 文, 30 名) を, 女性音声には ATR 音素バランス文 5118 発声 (503 文, 34 名) をそれぞれ使用した.

男声と女声の学習は別々に行った. そして, 得られた男声用 HMM(混合数 1) と女声用 HMM(混合数 1) の出力確率分布を混合することで, 混合数 2 の男女両用の HMM(男女混合分布 HMM) を作成した. 状態遷移確率は男女の平均値とした.

3.5 地名辞書の構造

地名辞書は図 3 のようにトライで表現した. これは合流のないオートマトンと考えることもできる.

トライの表現法の決定には計算量とメモリ量のトレードオフを考慮する必要がある. トライの操作に必要な計算量があまり多くならない範囲でトライデータの圧縮を行っている. 10,000 地名で約 140KW(1 ワード = 32bit) のデータ量である.

また, 「愛知県」の省略を可能にするために, 図中の太矢印が付いた 2 個所の状態を初期状態としている.

このとき, 県名, 市名, 町名などを単語と考えると単語 perplexity は 37.6 となる.

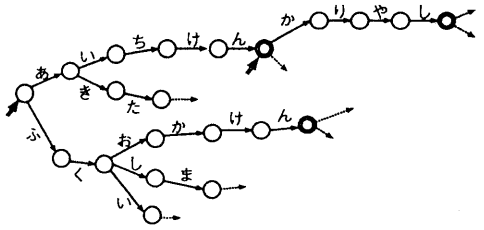


図 3 地名のトライ

3.6 認識アルゴリズム

特徴抽出部では, LPC 分析 (フレーム周期 8ms, フレーム長 16ms) によって, 16 次のケプストラム係数 (パワー項 c_0 を含む) と 16 次のデルタケプストラムを求めている. 音声区間検出に必要な零交叉周波数も計算する. 出力確率計算部では, すべての出力確率分布 (442 個) の出力確率を計算している.

認識は, オートマトン制御 One Pass DP 法 [3] にビームサーチによる枝刈りを組み合せたアルゴリズムで行っている. HMM の状態単位でビームサーチを行っている. ビーム幅は認識率の上昇がほぼ飽和する 1000 とした. 地名はトライによって表現しているため, 到達した状態と認識結果が一対一に対応している. そのため, バックトレース処理も, それに必要なデータも不用である. また, One Pass DP 法はフレーム同期なので, 認識

処理は他の特徴抽出や出力確率計算などの処理と並列に行っている.

すべての音声フレームの処理が終了すると, 最終状態 (図中の◎) の中から尤度の大きいもの n 個を認識結果として出力する.

音声区間の検出は, 外部から認識制御部を通じて伝えられる PTT(Push To Talk) 信号がオンになっている区間をベースに, それを音声パワーと零交叉周波数を使って延長や短縮することで行っている.

3.7 認識実験

地名認識の性能評価実験を行った. 学習データの話者とは異なる男性 11 名と女性 10 名が, 単語長 1～3 の地名をそれぞれ 50 地名ずつ計 150 地名を発声した音声を評価用データとした (男声 1650 サンプル, 女声 1500 サンプル).

正解の地名が 1 位として認識された確率と, 5 位以内で認識された確率を表 2 に示す. 表の 1 列目と 2 列目は, それぞれ男声用 HMM および女声用 HMM を使用したときの認識率である. 3 列目は, 3.4節で述べた男女混合分布 HMM を使用したときの認識率である.

表 2 認識結果

評価用音声	男声用 HMM	女声用 HMM	**男女混合分布 HMM**
男声	96.2% (99.4%)	78.6% (91.7%)	**95.7% (99.4%)**
女声	79.1% (90.5%)	92.8% (99.0%)	**92.2% (98.8%)**
平均	87.7% (95.0%)	85.7% (95.4%)	**94.0% (99.1%)**

() 内は 5 位以内認識率

男女混合分布 HMM を使用することで, 男声 (女声)HMM で男声 (女声) を認識したときと同等の認識率が得られることが分る. 男女の平均で, 1 位認識率は 94.0%, 5 位以内認識率は 99.1%となった.

4 地名認識の応用

4.1 音声入力公衆情報端末

前節で紹介した地名認識システムを使用した音声入力公衆情報端末について説明する (図 4).

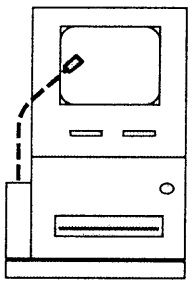


図 4 音声入力公衆情報端末

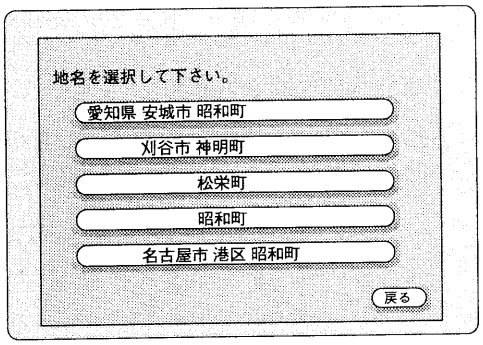
これのベースとなった公衆情報端末は, 我々が開発中の据置き型の情報端末であり, 高速道路のサービスエリアやガソリンスタンドなどに設置し, 地図の表示や経路案内, 渋滞や工事などの道路情報などの提供をするものである. 表示された情報のプリントアウトも可能である. 操作方法はタッチパネルによるメニュー選択方式である.

地図表示や経路案内などのサービスを利用するとき地名の入力が必要となる. この入力はタッチパネルによるメニュー選択で, 県・市・町を順に選択することで行っているが, 選択肢が多すぎて一画面に入りきらないことが多く, 操作に時間がかかっていた.

音声入力公衆情報端末は, 地名入力を音声でも行えるようにしたものである. このシステムでは, 画面上の PTT スイッチを押しながら, 例えば「愛知県刈谷市昭和町」と発声すると, 地名認識システムによって認識した上位 5 位までの地名が図 5(a) のように表示される. 目的の地名をタッチすると, 例えば図 5(b) のように地図が表示される.

候補の表示順は, ユーザが選択しやすいように, 認識順位順ではなく辞書式順としている.

音声入力とタッチパネルとを併用して地名を入力することもできる. また, 地名の音声入力は, 上記の例のように一度に行うことも, 県・市・町と一階層ごとに入力することもできる.



(a) 候補の選択画面

(b) 地図表示

図 5 画面表示例

4.2 評価実験

音声入力公衆情報端末における音声入力の有効性を調べるために, 26 名 (男性 23 名, 女性 3 名; 20 歳台 15 名, 30 歳台 8 名, 40 歳台 2 名, 50 歳台 1 名) に使用してもらい, アンケート調査を行った.

与えた課題は, 「愛知県知多市宝町」, 「愛知県刈谷市昭和町」や「自宅の住所」など 6 個所の地名を, 音声入力とタッチパネル入力とで交互に入力するものである.

課題終了後, 「タッチパネル入力は快適でしたか?」と「音声入力は快適でしたか?」という二つの問いに, 「たいへん不快」, 「不快」, 「やや不快」, ···, 「たいへん快適」という 7 段階で答えてもらった.

結果を図 6 に示す (一個の●が一人に対応している). これより, 多くの人が音声入力の方が快適であると感じていることが分かる.

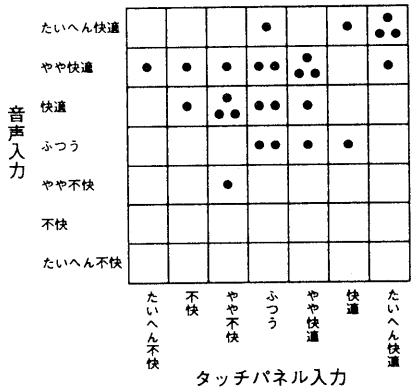


図 6 評価結果

5 まとめ

音声認識が役に立つようになるためには, 何が認識でき, 何が認識できないかがユーザにとって明確であることと, 語彙数が大きいこととが必要であるという考えに立ち, この二つを満足するタスクの具体例として地名認識システムを紹介した. これを応用した音声入力公衆情報端末の評価を行い, 地名認識の有効性を示した.

また音声認識を, 非常に多数のものから少数の候補を選択する手段とみなし, 最後の選択は他の方法で行う方法が有効であることも指摘した.

今後は, カーナビゲーションなど車載システムへの適用を進めて行きたい.

参考文献

[1] 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.

[2] Kai-Fu Lee: "Automatic Speech Recognition", Kluwer Academic Publishers, 1989.

[3] 中川聖一: "確率モデルによる音声認識", 電子情報通信学会, 1989.