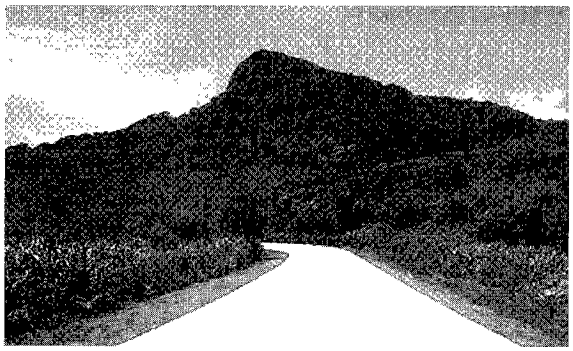
道しるべ：ここまできた音声認識技術



河原 達也
京都大学 情報学研究科

近年，音声認識システムの性能と実用化が著しく進展しており，今後さらにさまざまなアプリケーションへの展開が注目される．その際に最適な要素の選択やシステムの設計を行うためには，音声という信号やその情報処理に関する基本的な知識が必要不可欠である．本稿では，まず音声認識技術を種々の観点から分類し，その現状を概観する．次に，認識システムを構成する各要素に関して導入的な説明を行う．さらに，コーパスやツールキットなどのソフトウェア資源についても紹介する．

はじめに

自動音声認識というと一昔前は夢の技術の範疇であった．それが現在では，一般の人でもたいていは，パソコンソフトや携帯電話・カーナビなどで（少なくとも存在は）知っているのではないだろうか．率直に言って，研究者自身もこれほど早く展開すると思っていなかったふしもある．

実際に使った人の感想は，「ここまでできるなんてすごい」といった類と「あんなのまだ全然だめだよ」といった類に2極化されているようである．もちろん多くの技術に同様のことはいえるのだが，他のハードやソフト以上に，音声認識というのは，“**大変きまぐれなアナログ信号をごまかしのきかない文字テキストに置き換える**”という性質上，評価の揺れ（当りはずれ）が大きい．我々がデモをするたびに，ひやひやする由縁である．これらは，実際に話者による当たりはずれもあるし，単にそのときの周囲音の状況やマイクの微妙な位置による場合も多い．このように，**基本的な性能は着実に向上しているものの，音声認識技術はまだきわめて脆弱なものである．**

また一口に音声認識といっても，“Yes / No”の認識から任意語彙のディクテーションまで，多種多様である．研究者でもどこまでできているのか見解が分かれるものの，私の予想では，講演の書き起こしや会議の議事録の作成はいずれ（10年以内に）半自動でできるようになると思われる．一方，親しい人どうしが気軽にしゃべっているのを書き起こせるようになるのは，（それにどういう意義があるかは別にして）見当もつかないくらい先である．

このように，人間（ただし母国語話者）のような万能な認識能力を有するわけではないので，現状の音声認識技術を利用する際には，目的や条件に応じて最適な要素の選択や設定が必要である．実際に我々のシステムが他で使用される際にも，基本的な前提や要素選択が適切でないことがある．また，認識性能が十分に得られない場合も，ちょっとしたパラメータ設定の誤りによることが多い．本来はこうした煩わしさのない頑健なシステムの実現が我々研究者の責務であるが，音声認識を用いたシステムを設計・開発する方にとっても，エンドユーザにとっても，その中身を知っておくことは有意義であろう．そこで，現在どのレベルのことが可能であり，そのためにどのようなパーツが必要であるか，またそれを得るためのリンクについて紹介する．

音声認識の分類—できること—

音声認識技術をいくつかの観点から分類し，どういうことができてきたのかについて述べる．なお下記の記述には，分かりやすさのためと著者の主観のために，厳密でない点もあることをお断りしておく．

(1) 話者数による分類

1. 特定話者：ユーザを限定し，その人の発声を登録しておく．
2. 不特定話者：多数の話者で認識器を学習して，種々のユーザに対応できるようにしておく．
3. 話者適応型：最初は不特定話者であるが，ユーザの声に徐々に適応していく．

不特定話者の認識も可能になってきた．ただし小児の音声は別途の構築が必要である．話者適応を行うほうが望ましいが，ユーザにとって数十文もの事前発声登録（エンロールメント）は負担であるし，公衆サービスのようにそもそも適応が不可能な場合もある．

(2) 発声単位による分類

1. 単語認識：地名や人名など．「開けゴマ」のように一見文発声でも，文全体を登録しておけば単語認識と等価である．
2. 単語発声文認識：文を単語ごとに区切って発声する．初期のディクテーションがこの方式であった．
3. 連続音声認識：単語ごとに文を区切る必要がない（商品のカタログで書かれている）“自然な”発声．

単語認識はIC化されている．連続音声認識はアルゴリズムも複雑で，多大な計算資源を必要とするが，1990年代になって可能になった．

(3) 語彙サイズによる分類

1. 小語彙（数十単語）：数字認識，限られた人名認識（電話機など）／地名認識（券売機など）
2. 中語彙（数百単語）：ドメインを限定した案内などのサービス
3. 大語彙（5000単語以上）：全国の地名／人名（番号案内など），ディクテーション

大語彙の音声認識も可能になってきた．ただし，ICでできるのは中語彙程度である．

(4) 使用環境による分類

1. パソコン：入力の条件がよく，認識も比較的容易であるが，マウスやキーボードなど他の入力モードと競合する☆1．
2. 電話：帯域が限定され，回線中の雑音や歪みも大きいが，音声が主要なモードであるため，マーケットとしては有望である．
3. その他実環境：家の中，車の中，駅や店舗など．種々の雑音が激しい．

使用環境が厳しいほど，タスクを簡単にする必要がある．大語彙の連続音声認識が可能になってきたと述べたが，パソコンで接話型マイクを使用することが必要である．はじめに述べたように，周囲の雑音や残響条件によって大きく性能が劣化するのが，音声認識技術の最大の課題である．

(5) 発声スタイルによる分類

1. 読上げ音声（read）：ディクテーションにおいても，文章を推敲したうえで発声すれば，実質的には読上げ音声である．
2. 自発音声（spontaneous）：認識装置を意識するが，質問応答システムのように，思ったことをそのまま話す状態．
3. 会話音声（conversational）：認識装置を意識しないで，人間どうしで普通に話す状態．

現在の大語彙連続音声認識は，読上げ音声に近いことが前提である．自発音声を対象にする場合は，タスクや語彙などに制限を加えるのが普通である．会話音声の認識は，音響的にも言語的にも揺らぎが非常に大きいために現時点で未解決であるが，放送や講演など（認識装置でなく）公衆を意識して話されている音声を対象として，精力的に研究が行われている．

音声認識の要素技術—必要なもの—

音声認識のしくみ

典型的な連続音声認識システムの構成を**図-1**に示す．音声認識は，入力音声Xに対する事後確率$p(W|X)$が最大となる単語列Wをみつける問題として定式化され，これは音響モデルの確率$p(X|W)$と言語モデルの確率$p(W)$によって求められる．なお，単語認識は言語モデルがない場合に相当する．

音声分析は，音声のディジタル信号処理であり，主要な特徴量であるスペクトル包絡（あるいはその時間次元のケプストラム）を抽出するとともに，雑音除去や歪みの補

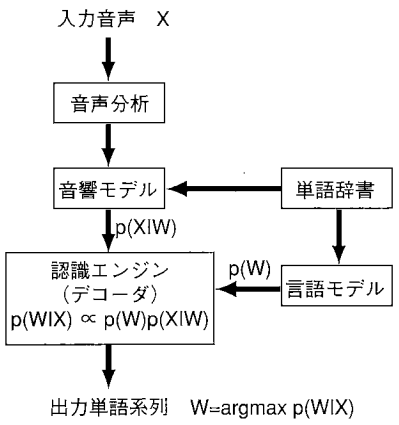


図-1 音声認識システムの構成

☆1 私の講義を受講した学生のレポート（複数）によると，「パソコンで音声認識を使いたいとは思わない」ということである．

正化も行う．

単語辞書は，認識対象の語彙エントリの発音（＝音素表記）を記述する．文字認識のように個々の文字を認識してから単語を認識するのではなく，認識対象の単語集合から可能な音素列を規定したうえで認識を行う．音声は，英語の続け文字や草書のように，個々の単位（＝音素）に区分化するのが容易でないためである．

音響モデルは，通常音素を単位とした音声特徴量パターンの分布の統計モデルであり，3状態程度のHMM（Hidden Markov Model）が主流である．同じ/k/という音でも，/k-a/の/k/と/k-i/の/k/では大きく異なるように，音素はその隣接音素に応じて細分化（＝コンテキスト依存）され，また多様な話者をカバーするために各状態で10以上の混合分布が用意される．その結果，音響モデルは合計数千から数万の正規分布で構成される．

言語モデルは，単語間の接続関係を規定する．オートマトンのような文法で書くと図-1のp(W)の値が0/1になる．機械翻訳と異なり，言語的に正しい入力を受理するだけでなく，正しくない仮説を生成しない能力が要求される．このような文法を記述するのは容易でないので，現実のテキストデータベースから単語の連鎖統計を抽出し，接続関係を確率値で与える統計的モデルが一般的になってきた．それも，単純な3単語連鎖（＝単語3-gram）モデルが主流である．ただし，数万の語彙ではその実存する組合せは膨大であり，数百万のオーダになる．

このように，**音声認識は膨大な音声パターンと言語パターンの統計データの集積**によって実現される．この膨大なデータを効率よく利用して，膨大な単語列の組合せから最適なものを探索するのが，認識エンジン，あるいは（発声された音声Xをテキスト Wに復号するという意味で）デコーダと呼ばれる．

アプリケーションやタスクに応じて適切な要素を選択する必要がある．以下にその指針を述べる．

音響モデルの選択

音響モデルは，その使用環境によって異なる．パソコンと電話と自動車内ではまったく異なるモデルと考えてよい．できるだけ同一の条件（帯域・入力装置・周囲音等）で収録された音声データで学習する必要がある．統計モデルである以上，その点が認識精度に決定的に影響する．また，タスクの複雑さによって必要な精度が異なり，大語彙ではコンテキスト依存モデルが必要である．さらに，発声スタイルによっても音声の特徴量パターンは異なるので，自発音声や会話音声では読上げ音声とは別のモデルを用意することが望ましい．

言語モデルの選択

言語モデルは，タスクが単純で明確な場合は文法を記述する．また，統計的モデルを学習するデータがなければ文法を書かざるを得ない．統計的モデルは学習が容易であるが，データを収集するのが大変である．このデータ収集も，使用される場面と同様の設定・条件にすることが重要である．たとえば音声対話システムでは，人間どうしでなく，(Wizard of Oz法☆2により見かけ上でも）機械を相手に発話してもらうことが望ましい．ディクテーション用のモデルは新聞記事等から学習されており，書き言葉の入力には動作しても，話し言葉には十分に対応できない．自発音声や会話音声には，そのためのモデル化・学習が必要である．なお日本語の場合は，分かち書きされていない文を単語に区切る形態素解析や漢字の読み付与の問題にも留意する必要がある．

認識エンジンの選択

認識エンジンは，言語モデルが文法ベースであるか統計的モデルであるかによって（モジュールかオプションが）異なる．単語認識の場合は簡易なものになる．また語彙サイズなどによって，必要な計算資源（=CPUパワーとメモリ量）が異なる．ただし，現在市販されているパソコンの能力であれば，大語彙でも十分である．

文献と技術情報へのリンク

教科書

これから音声認識の理論や手法を学ぶ方のために代表的な日本語の教科書を下記に挙げる．英語の著書も多数あるが，誌面の都合のため割愛する．

- 古井貞熙: 音声情報処理, 森北出版, 1998.
  ┬音声処理全般についてカバーされている
- 鹿野清宏他: 音声・音情報のディジタル信号処理, 昭晃堂, 1997.
  ┬音声認識に焦点をあてて丁寧に解説されている
- 中川聖一: 確率モデルによる音声認識, 電子情報通信学会（コロナ社）, 1988.
  ┬認識アルゴリズムが詳細に記述されている
- 中田和男: （改訂）音声, コロナ社, 1995.
- 今井　聖: 音声信号処理, 森北出版, 1996.
- 北, 中村, 永田: 音声言語処理, 森北出版, 1996.
- Rabiner, Juang （古井監訳）: 音声認識の基礎, NTTアドバンステクノロジ, 1995.

学会・研究会

音声認識技術に関して，研究者・開発者が会して情報交換を行うオープンな会議について紹介する．

発表件数・参加者が最も多いのは，日本音響学会の年2回の研究発表会で音声認識関連だけで毎回100件ほどの発表がある．研究会では，本会の音声言語情報処理（SLP）研究会をはじめとして，活発な発表・議論が行われている．

国際会議では，毎年開催されるIEEEのICASSP（Int'l Conf. Acoustics,Speech & Signal Processing）において，音声関連で200件超，うち認識関連で100件ほどの発表が行われる．また，隔年で交互に開催されるICSLP（Int'l Conf. Spoken Language Processing）とEUROSPEECH（ISCA主催）におい

☆2 機械が認識・応答しているようにみせかけて，ネットワーク経由などで人間が出力を生成する方法．

て，音声関連の発表が多数（700～800件程度）行われている．特にここ数年は世界的に音声認識技術の実用化への気運を背景として，これら以外のワークショップも含めて，いずれの会議も大盛況である．

- 情報処理学会 音声言語情報処理研究会（SLP）
  http://winnie.kuis.kyoto-u.ac.jp/sig-slp/
- 電子情報通信学会 音声研究会（SP）
  http://www.ieice.or.jp/iss/sp/jpn/sp-index-j.html
- 人工知能学会 言語・音声理解と対話処理研究会（SLUD）
  http://winnie.kuis.kyoto-u.ac.jp/sig-slud/
- 日本音響学会（ASJ） http://wwwsoc.nacsis.ac.jp/asj/
- IEEE Signal Processing (SP) Society
  http://www.ieee.org/organizations/society/sp/
- ISCA: Int'l Speech Communication Association
  http://www.isca-speech.org/
- 音声関連の文献データベース
  http://winnie.kuis.kyoto-u.ac.jp/bibliography/00README.html
- 音声・音関連のリンク集 http://www.aist-nara.ac.jp/IS/Shikano-lab/database/internet-resource/www-site.html

ソフトウェア資源

商品

パソコン上の日本語大語彙連続音声認識（ディクテーション）ソフトとして，IBMのViaVoice，NECのSmartVoice，Dragon Systemsの DragonSpeechが発売されている．Lernout & Hauspie (L&H)やPhilipsも，日本語版はまだであるが，英語をはじめとして各国語対応を進めている．それぞれSAPIに準拠したアプリケーション開発キット（SDK）も用意されている．

- IBM ViaVoice http://www.ibm.co.jp/voiceland/
- NEC SmartVoice http://www.psinfo.nec.co.jp/smart/
- Dragon NaturallySpeaking http://www.dragonsystems.com/products/
- L&H VoiceXpress http://www.lhs.com/voicexpress/
- Philips FreeSpeech http://www.speech.be.philips.com/
- Microsoft SAPI (Speech API)
  http://microsoft.com/iit/projects/sapisdk.htm
- 電子協 音声認識・合成関連製品動向調査結果
  http://www.jeida.or.jp/committee/humanmed/speech/

(フリー) ツールキット

主に研究者や開発者を対象とした，フリーあるいはそれに近いソフトウェアを紹介する．

IPA（情報処理振興事業協会）の補助で開発された「日本語ディクテーション基本ソフトウェア」は，オープンソースの認識エンジンJuliusと標準的な音響モデル・言語モデルなどのモジュールから構成され，汎用性・拡張性が高い．

統計的言語モデル（N-gram）作成にはCMU-Cambridgeツールキットが，音響モデル（HMM）作成ツールキットとしてはHTKが，それぞれ完成度が高く，「日本語ディクテーション基本ソフトウェア」もこれらを利用している☆3．

日本語の言語モデル化においては形態素解析システムが不可欠であるので，それらのリンクも付記しておく．

- 日本語ディクテーション基本ソフトウェア 認識エンジンJulius
  http://winnie.kuis.kyoto-u.ac.jp/pub/julius/
- CMU-Cambridge statistical language modeling toolkit
  http://www-svr.eng.cam.ac.uk/~prc14/toolkit.html
- Entropic HTK (HMM Toolkit)
  http://www.entropic.com/support/Patches/htk.html
- 形態素解析システム JUMAN
  http://pine.kuee.kyoto-u.ac.jp/nl-resource/juman.html
- 形態素解析システム 茶筌 (ChaSen)
  http://cl.aist-nara.ac.jp/lab/nlt/chasen.html

音声言語データベース

音声認識システムを構築するうえで必要不可欠となるのが，音声言語データベース・コーパスである．前述の通り，これらは目的に応じて収集する必要があるが，研究・開発のためにそれらの共有化が推進されている．

日本音響学会の音声データベース調査委員会では，標準的な読上げ音声データベースが作成された．また，ATRで収集された自然発話データベースも成果物として販売されている．その他のプロジェクト等においても，音声データベースの収集・公開がなされている．

このように種々のコーパスがあると，権利関係も含めて個別に照会・許諾を行う必要がある．そこで米国では，LDC (Linguistic Data Consortium) が一手に窓口となって，収集・頒布を行っている．日本でも同様の組織をめざして，言語資源共有機構（GSK）が発足している．

- 日本音響学会読み上げ音声コーパス（JNAS）
  http://www.milab.is.tsukuba.ac.jp/jnas/
- ATR音声言語データベース
  http://results.atr.co.jp/products/
- 電子協 音声データベース
  http://www.jeida.or.jp/committee/humanmed/speech/
- RWCP音声データベース
  http://www.rwcp.or.jp/wswg/rwcdb/speech/
- 文部省重点領域研究「音声対話」コーパス
  http://winnie.kuis.kyoto-u.ac.jp/taiwa-corpus/
- LDC: Linguistic Data Consortium http://www.ldc.upenn.edu/
- 言語資源共有機構（GSK） http://www.jeida.or.jp/gsk/
- 音声言語コーパスリンク集 http://www.milab.is.tsukuba.ac.jp/corpus/

不特定話者で十分な精度を得るには，数百人・計100時間以上の音声データを用いて音響モデルを学習する必要があり，数十GBもあるデータの統計処理には何カ月も要する．このような**大規模なデータによる高精度なモデルが現在の音声認識技術の基盤**になっている．

参考文献

1) 特集 音声処理技術とその応用，情報処理，Vol.38, No.11 (Nov. 1997).
2) 特集 音声言語情報処理の現状と研究課題，情報処理，Vol.36, No.11 (Nov. 1995).
3) 西村雅史，伊東伸泰：音声ワープロ，情報処理，Vol.40, No.2, pp.164-167 (Feb. 1999).

(平成12年3月3日受付)

☆3 ただしHTKは開発元のEntropic社が昨年末にMicrosoftに買収されたため，その取扱いが現段階で不透明である．