Techno Box

モバイルマルチメディア最新技術

その３　モバイル環境に向けた音声認識技術の現状

人間のコミュニケーションにおいて最も自然な手段である音声については，従来から活発な研究が行われてきた．中でも音声認識技術は近年急速に発展し，実用化されるようになってきている．本稿では，特にモバイル環境に着目し，音声認識技術の仕組みと課題，ならびに操作手段の視点から見た音声認識技術の特徴・課題について述べる．

張　志鵬（ちょう　しほう）

大辻 清太（おおつじ　きよたか）

杉村 利明（すぎむら　としあき）

1. まえがき

音声は，人と人との最も自然で効率的な情報交換手段である．人とパソコンとの情報交換においても，音声が使えれば，パソコンへの入力が極めて効率的になると期待される．音声認識技術は1950年から現在まで多くの研究者によって進められてきた．特に，統計的パターン認識理論[*1]，隠れマルコフモデル（後述）[*2]，確率言語モデル技術[*3]とコンピュータの処理能力の著しい向上により，音声ワープロ，放送への自動字幕化，など応用が展開しつつある．また，電話を用いて音声で情報を得る情報検索サービスも米国を中心に広がっている．今後，携帯電話，携帯情報端末（PDA：Personal Digital Assistant）のような携帯端末によるインタフェースが主流になってくると，音声入力の役割が一層大きくなると期待される．本稿では，音声認識の基本的な仕組みと課題を述べ，さらに操作手段として見た音声認識の特徴・課題について説明する．

2. 音声認識の基本

2.1 音声認識の原理

人が音声を生成するプロセスは，**図1**のように表すことができる．メッセージ情報源で意図したメッセージMが生成され，言語チャネルによって単語系列Wに変換される．単語系列Wは，調音チャネルによって音の系列Sとして実現される．音の系列Sは話者の口から放射され，音響入力信号としてマイクロホンに到達する．この過程が音響チャネルである．音響入力信号Aはマイクロホンによって電気信号に変換され，何らかの歪みを受けて，音声認識のXとして入力される．以上の各チャネルの確率は$P(W|M)$, $P(S|W)$, $P(A|S)$, $P(X|A)$の確率分布で表現される．$P(M)$はメッセージの事前分布である．

現在の音声認識システムでは，以上の過程を逆変換してXからMを回復することを目的としている．ベイズ決定理論に基づくパターン認識理論に従えば，$X=x_1, x_2, \cdots x_T$が与えられた時の事後確率$P(M|X)$を最大にする単語列$W=w_1, w_2, \cdots w_T$列を選択する．この選択プロセスで認識される単語列Wは，$P(W|X)=P(X|W)P(W)/P(X)$と表すことができる．

ここで，条件付確率$P(X|W)$は，単語Wを音素などの基本単位の連鎖で表した音響モデルから特徴パラメータ列が出現する確率として計算され，単語列Wが発声される事前確率$P(W)$は言語モデル（文法）によって計算される．通常，音響特徴パラメータは，ケプストラム[*4][1]などが用い

＊1　統計的パターン認識理論：大量サンプル数に基づく統計的手法を用いたパターン認識
＊2　隠れマルコフモデル：時系列信号の確率モデル
＊3　確率言語モデル技術：単語列の出現確率で表す言語モデル
＊4　ケプストラム：ケプストラムは対数スペクトルのIDFTで定義され，スペクトラム包絡と微細構造を近似的に分離して抽出できるという特徴がある

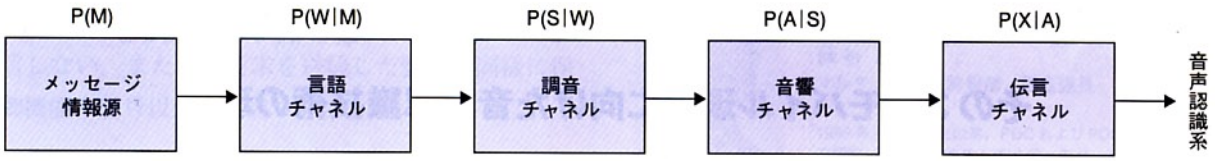

図1　音声生成モデル

られる．ここで，音響モデルとしては隠れマルコフモデル（HMM：Hidden Markov Model）[2]が用いられ，言語モデルとしては統計言語モデル[3]が用いられる．

2.2 音声認識の主な技術

(1)　音響分析

人間は，音声を聞きとる際に，スペクトル分析を行っていると考えられている[4]．音声認識においても短時間スペクトルの分析が重要である．音声は，声帯による音源（有声音源，無声音源）の成分が，喉から口にかけての声道の形状によって調音されることで生成される．このため，音声の短時間スペクトルは，音源に対応する，周波数方向に細かく変化する成分（微細構造）と，声道の形状による調音に対応する，緩やかに変化する成分（スペクトル包絡）の積となる．

音声の認識において重要な音韻性の識別に必要な情報は，スペクトル包絡に集中している．したがって，短時間スペクトルからスペクトル包絡を抽出する方法が重要となる．これはノンパラメトリックな抽出法とパラメトリックな抽出法に大別される．前者には，フィルタバンクを用いる方法[5]やFFT（高速フーリエ変換），もしくはそれらに基づくケプストラム分析[6]などがある．後者には，線形予測分析法[4]やそれに基づくケプストラム分析法である線形予測係数（LPC：Linear Predictive Coefficient）ケプストラム分析[6]などがある．

(2)　隠れマルコフモデル

近年，音声パターンの生成を直接確率的にモデル化したHMMによる音声認識手法[2]が提案され，急速に音声認識の性能および機能が発展した．HMMは，出力シンボルによって一意に状態遷移先が決まらないという意味での非決定有限状態オートマトン*5として定義される．すなわち，出力シンボル系列が与えられても状態遷移系列は特定できない．観測できるのはシンボル系列だけであることからHidden（隠れ）マルコフモデルと呼ばれている．HMMは，非定常信号源である音声信号を定常信号源の連結で表す統計的信号源モデルであり，動的計画法*6（DP：Dynamic Programming）マッチングによる方法に比べて，スペクトル時系列の統計的変動をモデルのパラメータに反映させることができる特徴がある．HMMには，ある状態からすべての状態に遷移できる全遷移型（Ergodic）モデル[2]や，状態遷移が一定方向に進むleft-to-rightモデル[2]などがある．通常，音声認識では，音声パターンの時間的な不可逆性の性質のために，left-to-rightモデルが用いられる．**図2**で例を示す．

(3)　言語処理

音声認識を探索問題と見なして実行する場合，音響的

*5　オートマトン：入力と内部状態から出力と次の内部状態が自動的に決定されるシステム
*6　動的計画法：最適化問題をとくためのアルゴリズムとしてよく利用されている手法

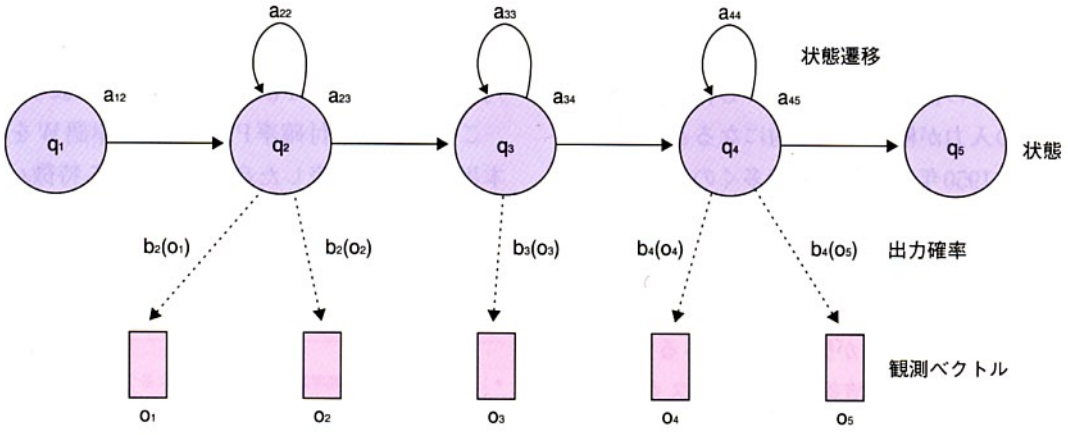

図2　HMM

な処理のみで認識を行い，後に言語的な解析を行う方法もあるが，一般には，探索に制約を設けて言語的な解析と音響的な解析を同時に行う．その制約を記述したのが言語モデルであり，代表的なものとして，2単語組モデル「bigram」，3単語組モデル「trigram」[3]がある．これは，単語・品詞の2, 3単語組確率を用いて単語列（文）の生起確率を近似するものである．2, 3単語組確率は訓練サンプルから統計的に求められるが，大量の訓練サンプルを必要とし，タスクで扱う語彙数が多い場合には，文の生起確率を精度よく近似することは難しいとされてきた．しかし，近年では大量の訓練サンプルが用意され，またそれを扱える計算機パワーが容易に得られるようになったことから，主にディクテーションシステム*7などにおいては本モデルが主流となっている．

3. 音声認識の課題

現在の音声認識では，例えば過去のニュース原稿によって学習した言語モデルを用いて放送音声を認識すると，スタジオで収録したアナウンサーの読上げ音声であれば90％以上の認識率が得られる．しかし，解説，現場の中継，対談などの音声に対しては大幅に認識率が下がる．この原因としては，原稿を読んでいない自発的な音声の言語構造と読上げ音声の違い，話者の音声および雑音などといった種々の音声変動が挙げられる．

3.1 話し言葉の音声認識の難しさ

自発的な話し言葉の音声認識の難しさには次のようなものがある．

*7 ディクテーションシステム：読み上げた言葉を書き起こすシステム

① 助詞などの言葉の脱落，省略，間投詞（あのう，えーなど）など不要語の付加
② 言い間違い，言いよどみ，言い直し，繰り返し
③ 言語モデルに含まれていない新しい言葉

これらの問題への対処法としては，大規模な話し言葉コーパス（大規模データベース）の構築，単語やフレーズを単位とした検出法，余計な言葉をスキップしながら文仮説とマッチングする方法などがある．

3.2 種々の音声変動

音声認識において，もう1つの大きな課題は，音声の変動を扱わねばならない点である．その背景には，話者の個人差が極めて大きいこと，種々の雑音とマイクロホンや伝送系の歪みなどが挙げられる．これらの音響変動に対応する耐性技術の向上が極めて重要である．**図3**に，主な音響的音声変動を示す．以下，主に雑音に対する適応手法について述べる．

実環境において音声認識を行う場合，雑音の影響が問題となる．走行中の車内環境の例では，エンジン音やロードノイズ，風切り音などのように，マイクから入り音声に影響を与える雑音や，マイクの違い，車内の形状，材質の違いなどによって生じる伝達特性に与える雑音がある．一般に，前者を加算性雑音と呼び，後者を乗算性雑音と呼ぶ．これらの雑音は，音声認識性能を著しく低下させる．このような雑音への対処法としてさまざまな方法が提案されている．

加算性雑音に対しては，SS（スペクトルサブトラクション）法[7]やHMM合成法[8]などのモデル適応法がある．SS法は，雑音が重畳された音声信号から，推定雑音をパワースペクトル領域上で差し引くことによって雑音を除去する方法である．雑音が定常に近い場合には非常に有効な手法であるが，非定常雑音への対処が困難である．HMM合成法は，種々の変動に対応できる基本的なモデル適応法として有望である．この手法では，音声のHMMと雑音のHMMの畳み込みをスペクトル空間で行って，雑音が重畳した音声のHMMを合成する．

一方，乗算性雑音に対し

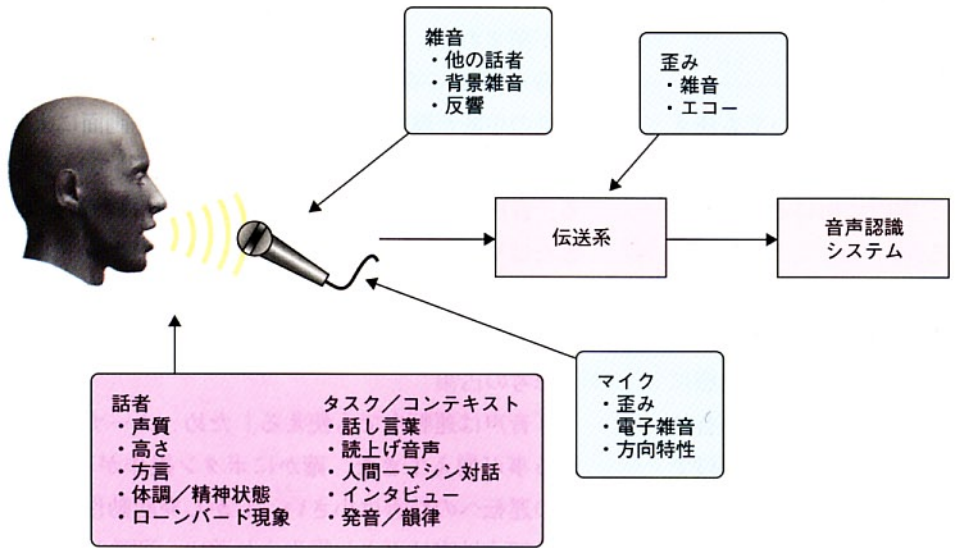

図3 音響的な変動

てはケプストラム平均正規化法（CMN：Cepstram Mean Normalization）[9]が一般的である．これは，音声をケプストラムで表現した場合，音声に対して乗算性雑音が加算的に現れることを利用し，ケプストラム領域で減算して補正するものである．この方法では，一般には発声が終了した時点でその発声に対する補正量を算出し，補正を施してから認識を行う．

4. 操作手段としての音声認識の特徴と課題

音声は，人間にとって自然な情報交換手段であるため，理想のインタフェースであると考えられてきた．上述のように近年の技術進歩により，原稿読上げ音声の認識は人間並みの性能を発揮するに至っている．しかし現在，音声認識がキーボード入力に取って代わる様子はなく，携帯電話サービスにおいても音声認識を使って大々的に成功した例はない．以下では，その原因と，それを踏まえたサービス適用の際の検討上の留意点について述べる．

4.1 音声認識と人工知能

音声認識に対して一般の人が抱く印象には，人工知能的要素がかなり含まれる．「話を分かってくれる能力」，つまり背景知識や状況・経緯についての理解，推論能力である．この能力は，人工知能的能力であり，文字による対話にも共通する課題である．これは高度な技術で，文字ベースでもまだ機械との自然な対話が実現しておらず，実現には相当な時間を要する．人工知能的能力の技術開発が進展するまでは，音声認識のサービスへの適用はタスクを限定するなどして，この能力がそれほどシビアに要求されず，擬似的に模擬できる対象に限定する必要がある．

4.2 音声の操作性能

音声は「訓練不要」「目や両手がふさがっていても使える」などの利点が強調され，特にモバイル環境に適した操作手段であると考えられてきたが，実際に用いるには，音声特有の性質を考慮する必要がある．

(1) オープンな伝達経路

音声は，発生元～受信先の信号伝達経路が空気であり露出している．このため，操作自体が周りの人に迷惑になる場所や操作内容を周囲の人に漏らしたくない場合には使えない．つまり，公共の場所では使いにくい手段である．また，受信された音の中から音声認識すべき対象を正確に取り出す技術が必要である．この点から，技術的要求が高く，操作の信頼度を上げにくい（実用上，間違いの許容やリカバリー，確認フローが必須となる）．現在，広く普及しているボタン操作の場合，伝達経路は指とボタンで直結し閉じており，どのボタンを押したかも利用者～機械間で明確であり，いずれの問題もない．

(2) 提示・閲覧・編集

音声だけを対話手段として用いた場合，提示・閲覧・編集は文字ベースのものよりかなり劣る．音声では複数の項目を一覧表示する事ができず，順次読上げとなる．よって利用者が入力すべき対象を正確に覚えていない場合には，推測や対話フローを工夫しないとかなり操作性が悪くなる．ある程度の長さや広がりを持つ文章，表などに対する閲覧・編集も，画面表示テキストの編集に比べ該当個所の指定・操作において劣り，これらが重要となるサービスには向かない．音声単独に適するサービスは，入力されうる対象が大量にあっても，利用者が正確に覚えている場合，例えば住所入力や着メロの曲名入力であろう．一般用途にはテキストなど他の入出力手段と併用する形態（マルチモーダル）での利用が妥当である．

(3) 保存・再利用

音声は聞けばすぐ分かる短い内容には適するが，ある程度長く記録を残したい場合には使いにくい．音声を残す「録音」という機能は，テキストに比べ必要な資源・操作性において劣る．また得た情報を一部引用するなど他に転用するのも難しい．聞き取った内容をメモ書きするぐらいなら，テキストでもらう方が手間がかからない．これらは音声を情報提供サービスに用いるときに評価・検討が必要な項目である．

(4) 一時中断

モバイル環境では，常に優先される別のタスク，例えば歩行や障害物回避などがあり，随時操作を中断できる事が望まれる．現在，一般的なのは，センターへ電話して音声回線にて音声認識するタイプである．この場合，対話の中断が難しい．iモードのボタン～画面による操作は，随時中断可能であるという点で適しているといえる．音声認識は，極めて短時間の対話で終わるタスクか，またはVoIP（Voice over IP）ベースでパケット通信などを用い，コストが時間によらず，かつ一時中断が可能なタスクに限定する必要があるだろう．

(5) 思考の占領

「音声は運転中でも使える」ため，カーナビに搭載される事が増えてきた．確かにボタン操作が不要であり，車の運転への妨害は小さい．しかし発話動作自体が思考に及ぼす妨害はボタン操作より強い．習熟したボタン操作は別のことを考えながらでも可能であるのに対し，音

声コマンドを発話しながら別のことを考えることは困難である．人間の思考は言語に依存する部分が大きいからである．この問題の一部は，前述した「一時中断」が容易に可能となれば解決できるが，この点がサービス設計上，大きな問題にならないかを検討しておくべきである．

以上のように，音声による操作は他の手段に対して優れた面があるものの，上述するような本質的な問題も持っている．よって，現時点では，それほど対話理解を必要としなくとも，サービスが組み上げられるものだけに適応していく工夫が必要だろう．

5. あとがき

音声認識システムには多様な応用が期待できるが，誰の声，どんな話題でも，またどんな環境でも，自然な話し方での音声を認識できるためには，解決しなければならない課題が多い．主たる課題は，話し言葉特有の言語，音響的変動，周囲雑音，電話音声の歪み，話者の個人差などに対するロバストなモデルと適応手法の構築である．その基礎となる大規模なコーパスや，実環境の大量音声データの構築も重要である．今後は単に音声から文字に変換するだけでなく，発声者の意図の理解，内容の自動要約へと展開し，さらに幅広い応用が期待される．

コンピュータの小型化，高性能化によりユビキタスコンピューティングの時代が到来すると言われている．音声は，上述したような長所と短所を持ちつつも，携帯電話の普及もあわせて，ヒューマンインターフェースの基本手段として広く使われるようになる可能性がある．その際，音声認識の多くは個人が身につけ，個人に特化したマシンで行われるようになり，そのシステム構成はこれまでと大きく異なってくるであろう．

文　献

[1] 板倉，東倉：“音声の特徴抽出と情報圧縮，情報処理”19，7，pp.644-656.

[2] J.R.Rabiner and B.H.Juang: “Fundamentals of Speech Recognition”. Prentice Hall, New Jersey, 1993.

[3] 中川聖一：“確率モデルによる音声認識”，コロナ社，1989.

[4] 板倉文忠，斎藤収三：“最尤スペクトル推定法を用いた音声情報圧縮”，日本音響学会誌，Vol.27，No.9，pp.17-26，1971.

[5] http://htk.eng.cam.ac.uk/

[6] 古井貞熙：“音響音声工学”，近代科学社，1992.

[7] S.Boll：“Suppression of Acoustic Noise in Speech Using Spectral Subtraction”, IEEE Transactions on Speech and Audio Processing, ASSP-27 (2), pp.113-120, 1979.

[8] F. Martin, et al：“Recognition of noisy speech by composition of hidden Markov models”, Proc. Eurospeech, pp.1031-1034, 1993.

[9] S. Furui：“Ceptral Analysis Technique For Automatic Speaker Verification”, IEEE Transaction on Acoustical Speech and Signal Processing, Vol.ASSP-29, pp.254-272, 1980.

用語一覧

APN：Access Point Name（接続ポイント名）
CiRCUS：treasure Casket of i-mode service, high Reliability platform for CUStomer
CMN：Cepstram Mean Normalization（ケプストラム平均正規化法）
DP：Dynamic Programming（動的計画法）
HMM：Hidden Markov Model（隠れマルコフモデル）
FOMA：Freedom Of Mobile multimedia Access
HTTPS：HyperText Transfer Protocol over SSL
ISP：Internet Service Provider
LPC：Linear Predictive Coefficient（線形予測係数）
MSISDN：Mobile Station international Integrated Services Digital Network number
PDA：Personal Digital Assistant（携帯情報端末）
PDC-P：PDC mobile Packet data communication system（PDC移動パケット通信システム）
RADIUS：Remote Authentication Dial In User Service
RFC：Request For Comments
TCP/IP：Transmission Control Protocol/Internet Protocol
VoIP：Voice over IP
W-TCP：TCP Profile over W-CDMA