# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題(和文) | 音声認識技術の最近の進歩 |
| Title | |
| 著者(和文) | 古井貞熙 |
| Authors | SADAOKI FURUI |
| 出典 / Citation | , Vol. 10, No. 3, pp. 251-254 |
| Citation(English) | , Vol. 10, No. 3, pp. 251-254 |
| 発行日 / Pub. date | 2000, 9 |
| 権利情報 / Copyright | 本著作物の著作権は日本応用数理学会に帰属します。<br>Copyright (c) 2000 Society for Industrial and Applied Mathematics. |

フォーラム

# 音声認識技術の最近の進歩

古井 貞熙

## 1 音声認識の原理

### 1.1 音声認識の基本的構成

音声認識(speech recognition)とは，音声波に含まれる意味内容に関する情報(言語情報)を，コンピュータや電気回路によって抽出し，判定することである．音声認識過程は，図1のように，通信理論(情報処理論)の問題として，確率モデルを用いて定式化することができる[1]．話者が文を考える過程が文発生部で，これを通信理論の情報源に対応させる．音声認識システムを音声分析部と言語復号部に分ける．話者による発声部と音声分析部を合わせて，一つの音響チャンネルとしてモデル化し，これを歪み(雑音)のある通信路に対応させる．音声認識システムの主たる部分である言語復号部を復号部に対応させる．話者は，情報源に対応する文 $w$ を頭の中で組み立て，それに基づいて，その話者の発話習慣に従って音声波形を生成する．これには通常，話者の個人差，付加雑音，伝送歪みなどが重畳している．音声分析部は音声波形データの分析・変換を行って，短時間スペクトルなどの時系列データ(ベクトル系列) $y$ を出力する．言語復号部は $y$ から送信文の推定値として $\hat{w}$ を出力する．$\hat{w}$ は事後確率 $P(w|y)$ が最大になるように推定する．$P(w|y)$ を直接求めるのは，通常困難であるので，ベイズ則(Bayes theorem)によって，次式を満たすように推定する．

$$P(\hat{w}|y) = \max_{w} \frac{P(y|w)P(w)}{P(y)} \qquad (1)$$

ここで，$P(y)$ は $w$ に無関係であるので無視することができる．尤度 $P(y|w)$ は音響モデルによって得られ，文 $w$ が発生される事前確率 $P(w)$ は言語モデルによって得られる．従って，音声認識のポイントは，$P(y|w)$ と $P(w)$ をいかに計算するか，言い換えると音響モデルと言語モデルをいかに作るかにある．

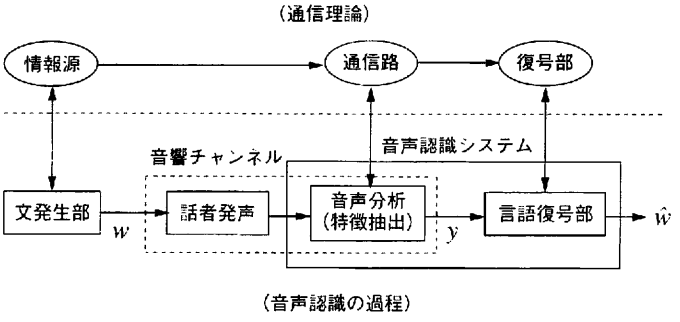


図1 音声認識過程の確率モデル

ふるい さだおき，東京工業大学大学院情報理工学研究科計算工学専攻．

### 1.2 音響モデル

基本単位としては音声の最も基本的な単位である音素を用いることが多いが，前後の音素の影響を考慮した複数のモデルを用いることが多い．このため，音素の数が数10程度であっても，直前あるいは直後のいずれかの音素を考慮する場合(このような音素単位をダイフォンと呼ぶ)で数100，前後両方を考慮する場合(このような音素単位をトライフォンと呼ぶ)には1000種類以上のモデルが必要になる．各 $w$ は音素の系列で表される．

尤度 $P(y|w)$ を計算するには，時系列データ $y$ の表現形式を決める必要がある．音声波形そのものを用いたのでは情報量が大きすぎ，波形の位相情報は伝送系や録音系によって変わりやすい上，人間による音声の知覚にはほとんど寄与しないので，音声波から一定周期ごとに，短時間スペクトル(密度)を抽出して用いる．現在では，スペクトルを対数変換した後，逆フーリエ変換して得られるケプストラム(cepstrum)を用いることが多い．短時間(対数)パワーも重要な特徴であるが，声の平均的大きさには意味がないので，数百ms～数sの平均パワーで正規化した値を用いる．

Δ(デルタ)ケプストラムとΔΔケプストラム[1]は，スペクトルの動的特徴(スペクトル変化)が音声知覚において重要な役割を果たしていることに着目して考案されたパラメータで，ケプストラムの時間変化の1次と2次の多項式展開係数である．伝送特性などの変動の影響を受けにくい特徴もある．対数パワーの展開係数とともに，元のケプストラムおよび正規化対数パワーと組み合わせて用いる．

## 2 HMM法

隠れマルコフモデル(Hidden Markov Model：HMM)は，上記の $P(y|w)$ を求めることを目的に考え出された方法で，各音素を確率状態遷移機械(マルコフモデル)で表現する[6]．非定常信号源である音声を，定常信号源の連結で表す統計的信号源モデルである．スペクトル時系列の統計的変動を表現できる特徴がある．音声認識に用いられるHMMは，left-to-right型で1つの初期状態と1つの最終状態がある構造が多く，図2は最もよく用いられるベイキス(Bakis)モデルと呼ばれる型の例である．図の状態遷移のアークに付けられた $a_{ij}$ は，状態 $q_i$ から状態 $q_j$ への状態遷移確率を表し，状態数を $S$ とすると $S\times S$ の行列で表現できる．通常，音声パターンには，時間的に非可逆性の性質があるので，$i>j$ なら $a_{ij}=0$ である．

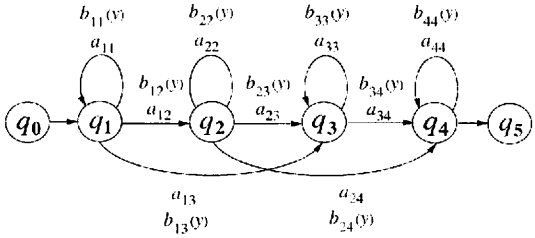


**図2 Bakis型HMMの例**
($q_i$：状態，$a_{ij}$：状態遷移行列，$b_{ij}(y)$：出現確率)

$b_{ij}(y)$ は，状態 $q_i$ から状態 $q_j$ への遷移で，スペクトルパターン $y$ が観測(出力)される観測(出現)確率を表し，$\{b_{ij}(y)\}$ を出現確率行列と呼ぶ．出現確率が状態遷移に独立に遷移元の状態によってのみ決定されるモデルも考えられ，この場合は $b_i(y)$ と書くことができる．この両者は数学的に等価変換が可能である．出現するスペクトルパターンに関しては，連続値として表す場合(連続分布モデル，連続HMM)と，有限個のシンボルの組み合わせで表現する場合(離散分布モデル，離散HMM)がある．連続分布モデルで出現確率を表す方法としては，単一ガウス分析(正規分布)または混合ガウス分布が用いられ，パラメータの自由度を減らすために無相関ガウス分布を用いることが多い．

HMMのパラメータは，バウムウェルチ(Baum-Welch)アルゴリズム(フォワードバックワード(forward-backward)アルゴリズム)によって，推定することができる．これは，適当に与

えた初期パラメータによるHMMを用いて，与えられた学習用サンプルに対する状態間の遷移確率を計算し，この遷移確率を確率的回数と仮定して，学習用サンプルに対する出現確率の最尤推定を行い，初期確率，遷移確率，出現確率の値を更新する手続きになっている．すべての学習用サンプルに関してこの計算を行ってから1回パラメータを更新するというサイクルを，値が収束するまで繰り返して，パラメータの値を決定する．EM (expectation-maximization)アルゴリズムの一種であり，少なくとも局所的最大値に収束する．

## 3 統計的言語モデル

言語ソースから生成させる単語列 $w$ の生起確率 $P(w)$ は，最近では統計的言語モデル[5]によって計算することが多い．与えられた単語列

$$w = w_1 w_2 \cdots w_k \tag{2}$$

の生起確率は，次のように表される．

$$P(w) = \prod_{i=1}^{k} P(w_i | w_1 w_2 \cdots w_{i-1}) \tag{3}$$

大語彙の音声認識では，長い系列について有効な統計量を得ることは極めて難しいため，マルコフ過程を導入し，単語の生起確率が直前の $N$-1個の単語のみに依存するとした $N$ グラム($N$-gram)が一般に用いられる．$N=2$ の場合がマルコフモデル(バイグラム; bigram)，$N=3$ の場合が2重マルコフモデル(トライグラム; trigram)である．さらに，$N$ グラムモデルの推定の不確実性を克服し，$N$-1グラム以下の比較的安定な統計量を用いて $N$ グラムを推定するための一手法として補間推定法(interpolated estimation)が用いられる．学習テキストに全く出現しない $N$ グラムの値を，$N$-1グラムの値から推定する方法として，グッドチューリング(Good-Turing)の推定法を基礎とするKatzのバックオフ平滑化法(backoff smoothing)が広く用いられている．

日本語のテキストから統計的言語モデルを学習する際の重要な問題は，英語などのテキストと違って，単語区切りを示すスペースなどがないことである．単語の定義すら必ずしもはっきりしていない．このため，単語ではなく形態素を単位と考えて，テキストを形態素解析し，形態素のバイグラムやトライグラムを用いるのが普通である[2]．

## 4 音声認識の今後の課題

音声認識の難しさ，すなわち誤認識の主たる原因は，音響モデルや言語モデルの学習に用いたデータと，実際の認識すべき音声の間に，必ず何らかのずれ(mismatch)があることにある[1]．その背景には，音声(声質や話し方)の個人差(方言や年令によるものも含む)が極めて大きいこと，同じ人でも体調や精神状態(ストレス)などによって発声速度や話し方が変化すること，多くの場合に，部屋の反響を含む種々の雑音が音声に加わっていて，しかもそれが時間的に変化すること，さらにその上にマイクロホンや電話機，伝送系などの歪みが加わることなどがある．これらの変動に対して認識性能が低下しない音声認識法，すなわちロバスト(robust)な音声認識法を構築することが，極めて重要である．種々の変動に対処するために，音声認識システムに，少量の音声サンプルに基づいて，変動に自動的に追随してくれるような，自動適応機能を持たせる研究が活発に行われている．

話し言葉特有のあいまいな文，文法的に誤った文，言い淀み，言い直し，言い間違い，省略，間投詞などを含む文をどのように言語モデルに反映させるかは，極めて難しい問題である．このような意味で，音声認識の本質を変えるアプローチとして，音声を忠実に文字に変換しようとするのではなく，発声者の意図や話題語(キーワード，キーフレーズ)を抽出することを目的とする枠組みや，音声を自動的に要約する試みも始まっている[3], [4]．

## 参考文献

[1] 古井貞熙，音声情報処理，森北出版，東京，1998.
[2] 古井貞熙，大語彙連続音声認識の現状と展望，日本音響学会春季研究発表会講演論文集，1-6-10, 1998.
[3] 古井貞熙，話し言葉の音声理解へ—存在感のある研

究を期待して一，電子情報通信学会技術報告，SP98-113, 1998.
[4] 堀智織，古井貞熙，話題語と言語モデルを用いた音声自動要約法の検討，電子情報通信学会技術報告，SP 99-110, 1999.
[5] 北研二，中村哲，永田昌明，音声言語処理，森北出版，東京，1996.
[6] Rabiner, L. R., and Juang, B.-H., Fundamentals of Speech Recognition, Prentice Hall, New Jersey, 1993.