論文 / 著書情報
Article / Book Information

論題(和文)	誰でも使える連続音声認識
Title	
著者(和文)	古井貞熙
Author	SADAOKI FURUI
出典(和文)	日本音響学会 第9回特別企画：ここまできた音声によるヒューマンインタフェース, Vol. , No. , pp. 1-10
Journal/Book name	, Vol. , No. , pp. 1-10
発行日 / Issue date	1996, 3

誰でも使える連続音声認識
Continuous Speech Recognition for Everyone's Use

古井　貞熙
Sadaoki Furui

NTTヒューマンインタフェース研究所
NTT Human Interface Laboratories
東京工業大学情報理工学研究科
Graduate School of Information Science and Engineering
Tokyo Institute of Technology
furui@splab.hil.ntt.jp, furui@cs.titech.ac.jp

1．まえがき

音声認識とは、人が発声した音声をコンピュータで知識処理することである [1] [2]。音声には、発声者が意図した言葉の意味内容の他、誰が話しているかという話者情報も含まれている。どちらも我々の日常の対話において、重要な役割を果たしている。音声による対話は、人と人との最も自然で、容易かつ効率的な情報交換手段である。人とコンピュータの間でも音声を使って対話ができるようになれば、極めて便利になると期待される。

音声認識技術の研究は、1950年代から今日まで、世界中の多数の研究者によって、40年以上にわたって行なわれてきた [1] [2] [3]。著者自身がこの研究分野に足を踏み入れてからも、すでに四半世紀が経過している。その間、Stanley Kubrickの有名な映画「2001年宇宙の旅」の中のコンピュータHALや、George Lucasの「スターウォーズ」のシリーズの映画の中のロボットC3P0を始めとして、音声認識のできる数多くのロボットや装置が、空想科学の中で人々の関心を集めてきた。話し言葉を認識し、その意味を理解し、誰が話しているかを判定することができる知的機械を作りたいという願望は、多くの研究者の夢であり続けてきた。

音声は目で見ることができないので、音声認識の難しさを直感的に把握するのは難しいが、いわば毛筆の続け字で書いた行書あるいは草書をコンピュータで読み取ったり、誰の筆跡であるかを判定するようなものである [3]。個人差や雑音（よごれ）によるパターンの変形が大きく、書き間違いに相当する言い間違いや言い直しも、日常の会話では頻繁に起こる。これらの問題にどの程度対応できるかが、重要なポイントである。これまでの40年間の技術的進歩により、音声認識のできるコンピュータの夢はやっと部分的に実現されつつあるが、任意の話題についての、あらゆる人による、あらゆる環境での音声を理解できるという目標からは、まだ程遠い状況にある。

コンピュータに、人が話した言葉を理解させる技術が狭義の音声認識技術、誰が話しているかを判断する技術が話者認識技術である。本解説では、スペースの関係から前者のみを対象とする。話者認識技術の動向については文献 [4] [5] を参照頂きたい。

2．音声認識の分類と基本的方法

音声認識は、次のように分類することができる。

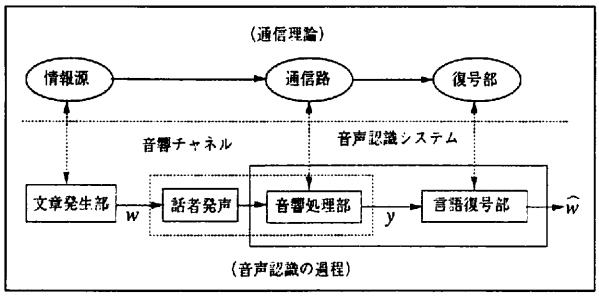
(1) （孤立）単語音声認識：区切って発声した単語を認識する。

(2) 単語スポッティング：文音声中のキーワードだけを認識する。

(3) 連続音声認識：単語を連続して発声した音

声を認識する。

連続音声認識はさらに、比較的多数の語彙を対象とし、文法などの言語的知識を用いて、その意味内容を理解しようとする文音声認識と、比較的少数の語彙を対象とし、単語間の言語的知識を用いずに、音響的特徴のみによって認識する連続単語音声認識に分けることができる。文音声認識は、会話音声認識あるいは音声理解ともよばれる。



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

図1　確率モデルに基づく音声認識システムのモデル化

音声認識システムは、図1に示すように、音声分析部とそれに続く言語復号(処理)部から構成し、情報処理論・通信理論の問題として、確率モデルを用いて定式化することができる [1] [2] [6]。この際、話者による発話と音声分析部を、一つの音響チャネルとしてモデル化する。話者は情報源に対応する文章wから、音声波を生成する。その音声波には通常、話者の個人差、付加雑音、伝送歪みなどが重畳している。音声分析部はスペクトル分析・特徴抽出を行なって、時系列データ(特徴パラメータ系列)yを得る。音響チャネルは雑音を含んだ通信路に対応し、言語復号部にyを送る。言語復号部はyから元の文章wを推定する。wは、ベイズ則によって展開した図中の式を最大化するように推定される。条件つき確率$P(y|w)$は音素などの音響モデルとの照合によって得られ、文章wが発声される事前確率$P(w)$は文法によって得られる。

人が音声を聞いて理解しようとするとき、音としての情報だけでなく、語彙に関する情報(どのような単語があるかという情報)を始め、種々の情報を組み合わせて音声を聞き取っている。多くの場合、状況や話題の展開から無意識に単語や句を予測したり、文脈から推し量って認識している。その一つの理由は、音素情報のあいまい性である。すなわち、通常の音声の中では各音素は必ずしもはっきりと発声されず、その部分だけを聞いたのでは何の音素か分からないことがめずらしくない。このため、コンピュータによる音声認識においても、語彙、文法などの情報を組み合わせて用いることが不可欠である [7]。

$P(w)$として、単語音声認識の場合は独立した単語の事前確率だけを、単語スポッティングの場合は文音声中にキーワードが存在する事前確率を、連続音声認識の場合は単語列の事前確率を用いる。英語のディクテーションシステムでは、単語毎に区切って発声した文章音声を認識する方式が用いられている。この場合は、音響処理は単語音声認識と同じに、$P(w)$は連続音声認識と同じになり、単語境界のセグメンテーションの必要がないだけ認識が容易になる。日本語では単語の定義が明確でないため、このような発声は難しい。

3．音声認識の基本技術

3．1　音声分析の方法

音声認識の第一歩は、音声の物理的特徴を、コンピュータで扱える形に変換することである。これを音声分析という。そのステップは次のようである [6] [8]。

(1) フィルタによる周波数帯域制限

次のステップの標本化において、標本化定理が満たされるように、音声波を周波数帯域を制限する低域通過フィルタに通す。4～8 kHz以下の周波数帯域に制限することが多い。

(2) 標本化、量子化

ディジタル処理ができるように、音声波を標本化し、さらに量子化する。

(3) スペクトル分析

音声波から、10ms程度の時間ごとに30ms程度の区間を切り出し、各区間から短時間スペクトルを計算する。スペクトル分析法としては、FFT(高速フーリエ変換)あるいは線形予測分析法を用いることが多い。

(4) 特徴パラメータ抽出

音声認識に最も重要な情報は、スペクトルの全体的な形状にある。これを安定かつ効率よく表現する方法として、通常、ケプストラム分析法を用いる。ケプストラムは、対数スペクトルの逆フーリエ変換である。さらに、スペクトルの時間変化に、聴覚的に重要な情報が含まれていることが分かっているため、ケプストラムの時間変化を表すΔケプストラムを抽出する。

(5) 音声区間検出

音声認識をするためには、音声が存在するか否かを検出することが必要である。この判定には、音声の振幅の大きさやその継続時間長などが用いられる。

3．2　音響モデルと音響処理

条件つき確率$P(y|w)$を求めるためには、音素などの音声の基本的な単位の標準パターンあるいはモデルを蓄えておき、認識すべき音声をそれらの単位と音響的に照合する。音声の基本単位としては、音素から単語まで種々のレベルが考えられる。音素は、音声の最も基本的な単位であり、日本語には、/a/、/i/、/u/、/e/、/o/の5種類の母音と、/p/、/t/、/k/などの約20種類の子音がある [6]。英語など、もっと多数の母音と子音をもつ言語も沢山あるが、音素の数が50を越える言語はほとんどない。すべての音声は、これらの音素の連続によって表すことができるので、音素を単位とすれば少数の基本単位と音声との照合によって音声認識ができそうに思えるが、実はそうではない。その主たる理由は調音結合、すなわち各音素の物理的特性が、前後の音素の影響によって変形する現象である [6]。

このため、同じ音素でも、その前後の音素に依存した複数のパターンあるいはモデルを用意する必要がある。音素の数が数十程度であっても、当該音素の直前あるいは直後のいずれかを考慮する場合（このような音素単位をダイフォンと呼ぶ）で数百、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）には千種類以上の単位が必要となる。図2に、トライフォンによって単語を表現した例を示す。

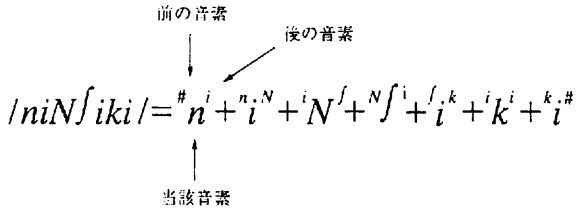


図2　「認識」という言葉をトライフォンの系列で表わした例
（#は無音を表わす）

単語に含まれる各音素の変形を忠実に表現するためには、各単語をそのまま単位として用いるのがよい。しかし認識すべき語彙の数が数千というように大きくなると、すべての語彙を蓄えるための記憶容量が膨大になるだけでなく、認識をするための計算量も膨大になってしまう。このため、語彙数が大きいときには、音素単位の結合として各単語を表現する方法を用いる。なお、音素と単語の中間である音節（日本語の「かな」に対応）や、それを半分に割った半音節と呼ばれる単位を用いる方法も研究されている [9]。

音声の特徴パラメータ系列と、各音素や単語の基本単位との類似の度合いを計算する際に最も重要な課題は、音声の時間的な伸縮にどう対処するかである。例えば、同じ音素でも、それが含まれる言葉、発声者、発声速度などによって長さが変化する。しかも、その変化はゴム紐のように線形（一様）ではなく、部分的に伸びたり縮んだりする非線形な伸縮である。このように非線形に伸縮するパターンを、そのずれを調整しながら照合する方法として、動的計画法（DP；dynamic programming）を用いた方法（DPマッチングと呼ばれる）が提案され、広く用いられている [6] [8]。

音声には、時間軸だけでなく、スペクトルの形そのものが、個人差を始めとする多様な要因によって変化する性質がある。この問題に対処するため、近年、隠れマルコフモデル（HMM；hidden Markov model）による方法が広く用いられている [6] [8]。図3に、HMMの例を示す。HMMは、確率状態遷移機械であり、各状態遷移確率と、状態遷移に伴う特徴パラメータの観測（出現）確率によって特徴付けら

れる。これにより、音声の確率的変動が、極めて効率よく表現できる。音声の各基本単位に対応するHMMをあらかじめ作成しておき、入力音声が各HMMから生成される確率（尤度）を計算する。この計算も、DPを用いることによって、能率よく行なうことができる。

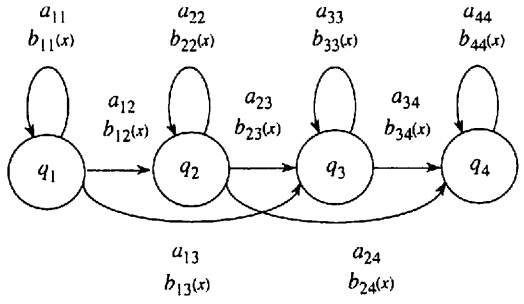


図3　HMM（隠れマルコフモデル）の例

（q_i：状態、a_{ij}：遷移確率、$b_i(x)$：スペクトルパターンxの出現確率）

認識すべき音声に対して、一つの基本単位を選択すればよい場合（典型的には、単語を単位として単語音声を認識する場合）は、これで条件つき確率$P(y|w)$が求まる。しかし、複数の基本単位の連続で音声が表現される場合には、あらゆる種類の基本単位の組み合わせ（並び）の候補と入力音声とを照合する必要がある。例えば7桁の数字の組み合わせは10^7の膨大な数となり、膨大な計算が必要となる。しかしこの場合も、DPを基本単位の中とその並びの2段階に用いることによって、大幅に計算量を減らすことができる。

数字の発音も、その前後の数字によって変化するので、それを考慮して各数字に複数の標準パターンあるいはモデルを用いる方法が研究され、これにより認識性能が大きく向上することが確認されている [11]。

3．3　言語モデルと言語処理

単語列の事前確率$P(w)$としては、単語の連鎖確率に代表される統計的言語モデル（stochastic language model）がよく用いられる [6] [7]。単語列

$$w_1^k = w_1 w_2 \cdots w_k \tag{1}$$

の生起確率

$$P(w_1^k) = \prod_{i=1}^{k} P(w_i | w_1 w_2 \cdots w_{i-1}) = \prod_{i=1}^{k} P(w_i | w_1^{i-1}) \tag{2}$$

は、$N(w_1^i)$を言語モデルの学習に用いるテキストデータベースに現れるw_1^iの頻度とすると

$$P(w_i | w_1^{i-1}) = \frac{N(w_1^i)}{N(w_1^{i-1})} \tag{3}$$

となる。ここで、単語数が2000種類で、i=4、すなわち4-グラム（4-gram）のモデルの場合を考えると、w_1^4の可能な異なり単語列の数は16兆（2000^4）通りとなり、有効な統計量を得ることは全く不可能である。統計的に可能なモデルとするために、マルコフ過程を導入して、マルコフモデル（バイグラム；bigram）や2重マルコフモデル（トライグラム；trigram）を用いる。バイグラムモデルでは、

$$P(w_i | w_1^{i-1}) = P(w_i | w_{i-1}) \tag{4}$$

と仮定し、トライグラムモデルでは、

$$P(w_i | w_1^{i-1}) = P(w_i | w_{i-2} w_{i-1}) \tag{5}$$

と仮定する。しかしながら、トライグラムモデルでの異なり単語列の数は80億（2000^3）通りとなり、まだ不可能に近い。一方、バイグラムモデルではとても構文的な情報を表わすことができない。

そこで、トライグラムモデルの推定の不確実性を克服するための一手法として考えられたのが、削除補間法（deleted interpolation）[6]である。削除補間法では、

$$P(w_i | w_{i-2} w_{i-1}) = \lambda_1 P(w_i | w_{i-2} w_{i-1}) + \lambda_2 P(w_i | w_{i-1}) + \lambda_3 P(w_i) \tag{6}$$

のように、下位の安定なバイグラムとユニグラム（unigram）の統計量を用いて補間することによって、トライグラムの確率を推定することが行なわれる。この削除補間法では、学習データの一部を削除してλ_1, λ_2, λ_3の推定を行ない、さらに、この削除する学習データの一

部を順次替えることを、学習データ全体で行なう。この推定では、補間の凸包性（convexity of interpolation）が保証されているため、λ_1, λ_2, λ_3の任意の初期値からアルゴリズムを開始することができる。

品詞（part of speech）のマルコフモデルを用いることも試みられている [7] [9]。また、単語でなく音素のバイグラムやトライグラムを用いる試みも行なわれている [10]。統計的言語モデルとしては、隣接関係だけでなくやや広い範囲の統計量として、単語の共起関係を用いる方法も用いられている [7]。これらの統計的言語モデルの有効な利用のためには、大量な言語データベースと、それからモデルを推定するための大量な演算が必要であるが、文書の電子化、コンピュータの高性能化などによって、これらが比較的容易に行なえるようになってきた。これにより、従来の人手に頼ることを前提としていた文法を用いるよりも認識性能が大きく向上したが、統計的言語モデルで扱える知識には限界がある。従来からの文脈自由文法、有限状態ネットワーク文法などの中に統計量を取り入れる試みも盛んに行なわれている。

文音声認識の研究の流れは、ディクテーションと対話音声の認識・理解を対象とした研究に分けられる。前者はいわゆる音声ワープロに相当し、書き言葉を発声したものを文字に変換するものである。そこで認識の対象とされる言葉は、書き言葉であるから、文法的にはほぼ正確であると想定することができる。そのかわり、認識しなければならない語彙や文体が極めて多様であるという点に難しさがある。後者の対話の場合には、対象（タスク）をある分野の情報案内などに限定すれば、語彙を制限することができるが、話し言葉特有のあいまいな文章や、文法的に誤った文章も対象としなければならない難しさがある。

4．実用化あるいは研究中の連続音声認識システムの例

単語音声認識装置は、すでにかなり実用化されている。世界で最初の大規模なシステムは、NTTが1981年に初めて実現した、音声認識を用いたバンキングサービスシステム「ANSERシステム」である [10]。最近では、電話の音声ダイヤルサービス、種々の情報案内、予約サービスなどが実用化されている。比較的最近実用化された大規模な音声認識システムに、米国AT&Tの電話オペレータの自動化システムがある [9]。このシステムが認識できる語彙は、「コレクトコール」、「指名通話」などの5種類のキーワードであるが、単語スポッティング法を用いているので、「お願いします」などの余分な言葉が付加してもよい特徴がある。音声に関する制約としては、その文音声の中に、キーワードが必ず一つ含まれているということだけである。

連続音声認識の内、連続数字音声認識はすでに実用化レベルにかなり近づいている。その用途は、音声ダイヤル、照会のためのクレジットカード番号の入力などである。

文音声認識はまだ研究段階にあるが、世界で最大のプロジェクトは、米国のARPA (Advanced Research Projects Agency) によるものである [9]。その内、ディクテーションについては、Wall Street Journal新聞の記事を読み上げた文章を認識するプロジェクトが強力に進められた。音素認識にはトライフォンのHMMを用い、文法には統計的言語モデルを用いて、65,000語の語彙を対象にした場合、93%の単語認識率が得られている。同じくARPAの対話音声の認識・理解プロジェクトとしては、ATISシステム (Air Travel Information System) が、やはり強力に進められてきている。これは、米国内の航空路線を用いて旅行をすることを想定して、コンピュータによる案内システムに、音声で問い合わせや予約をするシステムである [9] [12]。自由に発声した話し言葉を対象とするため、言い間違い、繰り返しなど難しい問題が多数含まれている。どこまで難しい音声を対象とするかによって異なるが、現在90～95%程度の単語認識率が得られている。

日本では、自動翻訳電話や電話番号案内システムを想定した研究などが行なわれている。最近では、Wall Street Journalに似ていると考えられる日経新聞の記事を読み上げた文章を認

識する研究も行なわれている [13]。約5年分の新聞記事680万文を用いて、頻度の高い7,000語を対象としたバイグラムに基づく言語モデルを作成し、トライフォンのHMMを用いた認識実験で、80%程度の単語（形態素）認識率が得られている。電話番号案内タスクでは、極めて多数の人の名前や住所を、かなり自由な文体で発声した音声の認識ができるように、約80,000語を語彙とし、文脈自由文法を用いた認識システムの実験が進められている [10] [14]。

現在の大語彙連続音声認識システムの典型的な構成を、図4に示す [1] [2]。大語彙連続音声認識では、ボトムアップに処理を行なったのでは、文章仮説（認識結果の候補）の数が爆発的に増えてしまう。このため、認識対象タスクに応じた言語モデル（単語辞書、意味情報、構文情報、文脈情報など）をトップダウンで用いて文章仮説を生成（予測）し、それを音素系列で表わしたものを逐次的に入力音声と照合して、検証していく方法をとる。

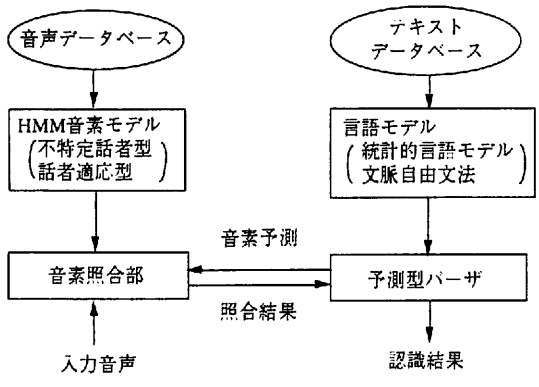


図4　現在の連続音声認識システムの典型的な構成

上述のATISシステムや電話番号案内システムでは、基本的にシステムへの入力は音声認識で行い、システムからの情報は画面上の表や文章で受け取る形をとっている。マルチモーダル、あるいはマルチメディア環境でのコンピュータとの情報の授受の方法として、これが人にとって最も容易かつ便利な方法と考えられるためである [10] [15]。

5．誰でも使える連続音声認識を目指して

5．1　音声認識の難しさ—予期できない変動—

「まえがき」でも述べたように、音声認識のできるコンピュータの夢はやっと部分的に実現されつつあるが、任意の話題についての、あらゆる人による、あらゆる環境での音声を理解できるという目標からは、まだ程遠い状況にある。その主たる理由は、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、必ず何らかのずれがあることである。その原因には、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おり、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 [15] [16] [17]。

最近の統計的方法により、これらの種々の音声の変動をカバーするような大量の音声データを用いてモデルを作成すれば、従来の方法に比べて高い認識精度を達成することができるようになった。個人差に対処するために、多数の話者の音声を集めて作成したモデルを用いて音声認識する方法は、「不特定話者型の音声認識」と呼ばれる。しかし、集められるデータ量には自ら限界があり、すべての変動条件を尽くすことは不可能である。さらに、データに含まれる変動があまりに大きいと、モデルがぼけて、異なる音素や単語の物理的特徴の重なりが大きくなるため、この方法には限界がある。このため、「sheep and goats 現象」と言われるように、大多数の話者に対してはうまく動作しても、少数ではあるが認識精度が極めて悪い話者や条件を生ずることになることが多い。

3.3節で述べたように、対話音声の認識・理解では、話し言葉特有のあいまいな文章や、文法的に誤った文章も対象としなければならない。この難しさも、話し言葉を対象とした文法が確立していないことと、その変動が予期できないことにあると言える [15]。

5．2　変動への自動適応機能

このため、音声認識システムに、少量の音声サンプルに基づいて、音声の種々の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ことが必要である [15] [16] [17]。人が音声を聴き取る際にも、新しい話者に対して、初めは何を言っているのかよくわからないが、5音節程度を聞けば、その声に慣れてよくわかるようになることが実験的に確認されている [17]。

一般に適応化アルゴリズムには、その規則を学習するためにあらかじめ決められた言葉を発声してもらう「教師つき（supervised）学習」と、任意の音声を用いて規則を学習する「教師なし（unsupervised）学習」による方法がある。前者の方が規則の獲得は比較的容易であるが、ユーザにとっては、認識装置を使う前にわざわざ特定の言葉を発声しなければならないのは煩わしい。後者の場合は、認識すべき音声をそのまま使って適応化（instantaneous adaptation）できるので、ユーザにとっては全く意識せずに、あたかも誰の声にも動作する装置であるように使えるので便利であるが、極めて個人差や歪みの大きい音声に対してこの機能を実現するのは、極めて難しい。適応化が真に必要なのは、このように認識率が特に低い話者や条件に対してであり、このときに効果がない適応化法は意味がない。このような難しさはあっても、実際に求められている技術として、教師なしで動作するアルゴリズムを今後追及していくべきであろう。さらに加えて、認識装置を使っているうちに、自動的に適応化が進んでいくこと（on-line incremental adaptation）が望ましい。

音声は通常、スペクトルあるいは対数スペクトル領域のパラメータで表現されるが、上で述べた種々の音声の変動は、これらの領域での線形あるいは非線形の変化として表される。さらに、その変化が、元の音声パラメータに対して、加算的な場合と乗算的な場合がある。また、その影響が種々の音素に共通に現れるものと、音素に依存して異なるものとがある。雑音やマイクロホンなどによる歪みは音素に共通であるが、個人差は音素によって異なる。この違いによって、適応化技術も異なってくる。また、これらの種々の変動は、単独にではなく複合して音声に加わるのが普通である。従って、これらに同時に対応できる適応化法を構築することが必要である。

もう一つ重要な技術的条件として、その適応化法が、基本的な音声認識アルゴリズムと良く整合することが必要である。例えば、これまでに述べたように、現在の音声認識法では、音声の特徴パラメータとしてケプストラムが用いられる。さらに全体的には統計的な枠組みを基本としており、尤度最大化基準でパラメータ推定が行われ、尤度の値に基づいて認識の判定が行われる。従って、適応化の方法も、これらに整合することが必要である。そうでないと、処理が複雑になるだけでなく、処理の最適性が保証されなくなる。

適応化法はさらに、音声の特徴パラメータの領域で適用する方法と、音響モデルの領域で適用する方法がある。前者は発声内容に無関係に適用できるので容易であるが、音素に依存した適応化ができない限界がある。

5．3　適応化技術

特徴パラメータ領域の代表的な適応化法に、ケプストラムの時間平均値を差し引く方法（CMN；cepstral mean normalization）がある [17]。数百ミリ秒あるいは文章の長さ程度にわたって、各ケプストラムの値を平均化し、その値を各時点のケプストラムから差し引く方法である。ケプストラム領域での引き算は、スペクトル領域での除算に相当するので、この処理により、平均的なスペクトル特性の逆フィルタが行われることになり、比較的定常的な歪みの影響を除去することができる。簡単ではあるが有効性の高い方法である。

音響モデルの領域で適用される基本的な適応化法に、ベイズ適応学習（Bayes adaptive learning）、あるいはそれから導かれるMAP推定（maximum a posteriori estimation）による方法がある [9] [17]。この方法は、不特定話者モデルのような、信頼性は高いが現在の入力音声には必ずしも合っていないモデルと、現在の入力音声に近いが少量のデータに基づいて

いるために信頼性が低いモデルを、適切に組み合わせて信頼性の高いモデルを作ることができる。

HMMを用いる認識系において、種々の変動に対応できる基本的な適応化法として有望なのが、HMM合成法 [18] である。図5に、その基本原理を示す。この方法では、雑音もHMMでモデル化し、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。このようにこの方法は、スペクトル領域での加算的な雑音に対処する方法として提案されたが、最近、図6に示すような、乗算的な歪みを同時に含む一般的な場合に適用できる方法に拡張された。図中、Gは発声方法の変化などを含む乗算的歪み、Hはマイクロホンや電話機、伝送系などの乗算的歪み、kは音声と加算的雑音のエネルギー比の変化に対応するパラメータを示す。実際の適応化アルゴリズムでは、尤度最大化の基準に基づいて、これらのパラメータを自動的に決定する。実験の結果、この方法の有効性が確認されている。

少量の音声サンプルを用いて適応化をする場合、この中には限られた数の音素しか含まれないので、含まれない音素を推定することが必要になる。この方法としては、スペクトル空間における近傍の音素の変化から補間（平滑化）する方法、変化に相関のある音素をあらかじめ統計的に調べておいて、この音素の変化から補間する方法、変化を比較的単純な線形変換（アフィン変換）や少数のバイアスに拘束する方法などが試みられている [17]。

従来の音声認識におけるパラメータ推定は、尤度最大化を基準にして行われることが多かった。しかし、音声認識の究極の目的は誤認識の最小化にあるので、この基準を直接的に用いてパラメータを推定する方法（MCE/GPD；minimum classification error/generalized probailistic descent algorithm）が、最近提案されている [17] [19]。話者適応化においてもこの基準を用いる方法が試みられ、有効性が確認されている [20]。

適応化が必要なのは音響処理のレベルだけでない。言語モデルに関しても、認識すべきタスク（認識対象）などが変わるたびに、大量の言語データベースを収集しなければならないのは、非現実的である。ユーザによっても、言語モデルは本来異なっていると考えられる。このため、できるだけ小規模のテキストデータに基づいて、いかにして言語モデルを新しいタスクやユーザに適応させるか、すなわち言語モデルをいかに効率的に自動獲得するかということが極めて重要である [17]。このためには、言語モデルが基本的に、データに基づいて自動的に適応化できる構造になっていることが必須である。

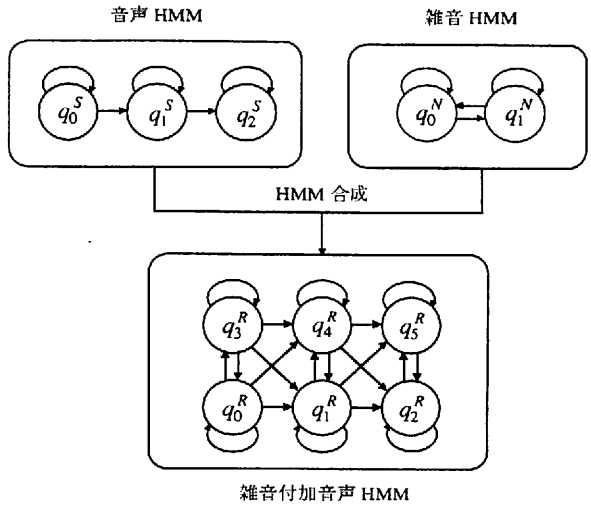


図5　HMM合成法の原理

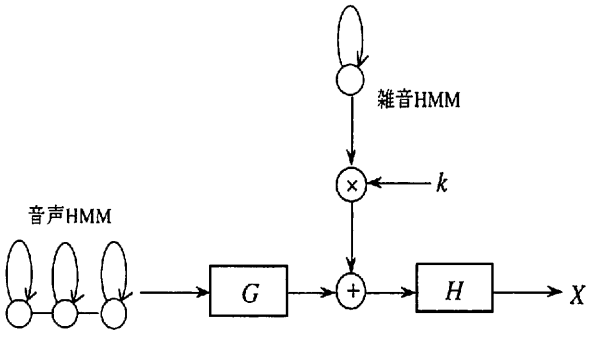


図6　雑音と歪みを含む音声のモデル化

5．4　音声の言語モデル

音声認識システムを実際に用いる際に、決められた言葉だけでなく、「えーと」「あのう」のような言葉（余剰語）が付加されるのが極めて自然な現象である。余計な雑音が入って来ることも多い。これらで認識システムが誤動作しないように、余計な音をいかにリジェクト、あるいは無視するかという機能も、システムを使いやすくするために必須である [15] [17]。さらに、1000語以上の大語彙のタスクになると、すべての認識対象語彙を覚えておくことは不可能なので、対象語彙以外の言葉（OOV；out-of-vocabulary words）で、しかも意味がある言葉を発声する可能性が大きくなる。この場合は、OOVであることを検出すると同時に、余剰語のように無視するのではなく、新しい言葉として適切に処理しなければならない。これは、極めて難しい課題の一つである。

人の自然な発声として、言い直しや繰り返しも頻繁に起こる。人とコンピュータとの対話においては、人と人の自然な会話におけるよりははるかに少なくなるが、無視することはできない。これらに対処するためには、すべての言葉を完全にとらえようとせず、むしろキーワードを手掛かりにして、文の意味をとらえるアルゴリズムが必要となると考えられる。

6．むすび

音声認識の最新技術の動向と、今後の展望について述べた。当面の実用化としては、単語音声認識や少数の語彙のスポッティングのような、基本的な技術によって実現可能な応用を探索していくのが現実的であるが、音声認識が広く用いられるためには、大語彙の連続音声認識技術を確立することが不可欠である。しかも、それが誰でもどこでもどのような対象に対しても確実に動作することが必要である。このために越えなければならないハードルは多いが、適応化技術を中心とする着実な研究開発努力によって、必ずやその目標は達成できると考える。

スペースの関係で触れることができなかった技術として、複数のマイクロホンを用いて、異なる方向の音源を分離し、マイクロホンから離れたところで発声した音声をとらえる方法がある [10]。これによって認識装置の使い勝手や性能を大きく向上させることができる可能性があり、米国のARPAによる音声認識のプロジェクトでも最近取り上げられている。また、本文で、マルチメディアやマルチモーダル環境での音声認識の利用について考えることの重要性に触れたが、さらに仮想現実（virtual reality）環境を考えることによって、さらに新たな技術の視点が浮かび上がってくることも期待される [15]。今後、これまでに増して、他の技術分野との融合や協力関係が求められてくると予想される。

参考文献

[1] 古井貞熙：“音声認識”、信学誌、Vol. 78、No. 11、pp. 1114-1118 (1995)

[2] 古井貞熙：“音声認識技術の現状”、情報の科学と技術、Vol. 46、No. 1 (1996)

[3] 古井貞熙：“音声認識の過去・現在・未来”、NTT R&D、Vol. 43、No. 10、pp. 1055-1060 (1994)

[4] 古井貞熙：“声の個人性の話”、日本音響学会誌、Vol. 51、No. 11、pp. 876-881 (1995)

[5] 松井、古井：“話者認識技術”、NTT R&D、Vol.43、No. 10、pp. 1101-1108 (1994)

[6] 古井貞熙：“音響・音声工学”、東京、近代科学社 (1992)

[7] 松岡、南：“音声認識の言語処理技術”、NTT R&D、Vol.43、No. 10、pp. 1091-1100 (1994)

[8] 南、松岡：“音声認識の音響処理技術”、NTT R&D、Vol.43、No. 10、pp. 1081-1090 (1994)

[9] C.-H. Lee、L. R. Rabiner：“音声認識の将来”、NTT R&D、Vol.43、No. 10、pp. 1069-1080 (1994)

[10] 鹿野清宏：“音声認識技術の現状と展望、NTT R&D、Vol.43、No. 10、pp. 1061-1068 (1994)

[11] 松岡、植本、松井、古井：“連続数字音声認識における音響モデル学習法の検討”、

信学技報、SP95-23 (1995)
[12] 松岡、Hasson、Barlow、古井："テキストコーパスを用いた音声理解のための言語モデル自動獲得"、信学技報、SP95-94 (1995)
[13] 大附、森、松岡、古井、白井："新聞記事を用いた大語彙連続音声認識の検討"、信学技報、SP95-90 (1995)
[14] 古井貞熙："10万語の連続音声認識アルゴリズム"、Computer Today、No. 59、pp.24-28 (1994)
[15] S. Furui : "Prospects for spoken dialogue systems in a multimedia environment", Proc. ESCA Workshop on Spoken Dialogue Systems, Vigso, Denmark, pp. 9-16 (1995)
[16] S. Furui : "Recent advances in speech processing technology", Proc. 15th International Congress on Acoustics, Trondheim, Norway, Vol. III, pp. 497-504 (1995)
[17] S. Furui : "Flexible speech recognition",Proc. Eurospeech'95, Madrid, Spain, pp.1595-1603 (1995)
[18] 南、古井："HMM合成に基づく尤度最大化適応法"、信学技報、SP95-2 (1995)
[19] S. Katagiri, C.-H. Lee and B.-H. Juang : "New discriminative algorithm based on the generalized probabilistic descent method," Proc. IEEE Workshop on Neural Networks forSignal Processing, Princeton, pp. 299-309 (1992)
[20] T. Matsui and S. Furui : "A study of speaker adaptation based on minimum classification error training", Eurospeech'95, Madrid, Spain, pp.81-84 (1995)