音声認識技術の研究

名古屋大学情報科学研究科 教授
武田一哉

1 音声認識とは

インターネット，携帯電話，大容量記憶，高速コンピュータ，など，電子情報通信技術は，急速な発展を続けています。このような社会技術基盤の発展は，我々の生活に2 つの革新的な変化をもたらしつつあります。第一の変化は，「遍在化」すなわち，いつでもどこでも高度なネットワークを利用して，コミュニケーションすること，多様なサービスにアクセスすることが可能となったことです。今ではすっかり街で見慣れた光景になりましたが，携帯電話のボタンを操って，メールをやりとりすることは，つい5年前では珍しいことでした。第二の変化は，サービスの高度化です。チケットの予約や，通信販売など高い利便性を享受することが可能になった反面，その複雑な操作に戸惑う利用者も少なくありません。

情報サービスの遍在化と高度化に伴い，システムの操作性が問題となりつつあります。外出先で，キーボードやディスプレイを使わずに，複雑なサービスにアクセスする手段が求められています。人に対するのと同じように，情報システムに対しても，音声による対話で操作を行うことが望まれます。音声認識はこのように，音声を自動認識する機能です。多くの音声認識システムには，単に話した音声を文字に変換するだけでなく，その意味を（文脈を考慮にいれて）正しく理解する，ことが求められます。しかし「人間なみ」の理解力を持った音声認識システムを実現することは困難なことが現状です。本稿では，音声認識の仕組み，研究の動向，音声認識の応用と展望，について，解説します。

2 音声認識のしくみ

2.1 音声の生成

音声認識の原理を解説する前に，人間が音声を生成する一般的なしくみについて概説したいと思います。音声すなわち言語音は，声帯の振動によって作られた音に，声道の形状によって定まる音響的な伝達特性によって，音色が付加されることで生成されます。声帯の振動は，その周期が短いほど，高い声として発声され，周期が長いほど，低い声として発声されます。また声の大きさも，声帯の振動の振幅によっておおよそ決まります。

一方声道の形は，舌の位置や顎の構えによって時刻とともに様々に変化し，それに応じた音色が声帯振動に付加されるのですが，その音色こそが言語音を特徴付けています。同じ言葉を発声する時の口の形や舌の位置は，いつもおおよそ同じであるはずです。このように，声道の形状によって定まる言語音の物理的な性質から，発声された言葉を推定することが音声認識の基本的な問題です。

言語学的には，音声は音素と呼ばれる基本単位に分割することができます。例えば，「日本語は」という分節は／n, i, q, p, o, N, g, o, w, a／ のように10個の音素から構成されます。（ここではqで促音を表しています。）しかしこのことは音声の物理的な波形が，10の区間に時間的に分離できるということではなく，音声波形上では音素と音素の境界は明確でありません。このことは，図1に示した音声の時間波形と声紋からも見て取れると思います。このように，音素と音素の境界が明確でないことは，音声認識を困難にしている大きな要因の一つです。

2.2 音声の特徴分析

音素に関する情報が，声の高さとは独立した情報であることは，同じ単語を異なる声の高さで発声しても，その言語的な意味に変わりがないことからも想像できると思います。（但し，橋と箸のように，アクセントを変えると意味が変わってしまう場合も，まれには存在しますが。）多くの日本語音声認識システム

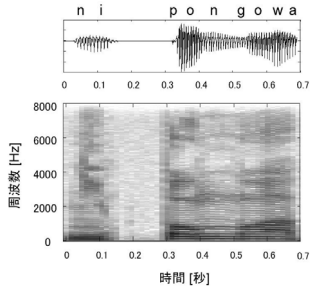

図1　音声の時間波形と声紋

では，声の高さや強さは，音素の情報とは無関係な情報として，音声から取り除かれます。具体的には，入力音声の基本周波数成分を除去する，ケプストラム（Cepstrum）分析(*1)が利用されることが一般的です。ケプストラム分析を行うことで，声の高さや声の大きさといった情報を音声から取り除き，音素に関係する（声道の形状に関係する）情報のみを取り出し，10～20次元程度のベクトルの形式で表すことができます。

ケプストラム分析は，1秒間に100回すなわち10ミリ秒程度の間隔で行われ，時間とともに連続的に変化する信号が，音声認識に利用されます。

2.3 音声のモデル化

分析の結果得られるケプストラムの系列は，声道の形状に対応しています。そこで，音素毎にケプストラムの典型的な系列を記録し，入力された未知の音声から得られたケプストラムの系列と，この典型的な系列とを比較することで，最も基本的な入力音声の認識が実現されます。しかし，実際にはここで2つの問題があります。すなわち（1）入力された音声は音素毎に区切られていないこと。（2）予め蓄えられている，標準的なケプストラムの系列の長さと，入力される音声中の音素の長さは，一般には一致しないこと。の二点です。

そこで，同じ音素に対して，異なるケプストラムの系列を出力しうるような，確率的なモデルとして，現在は隠れマルコフモデルが広く利用されています。隠れマルコフモデルは，確率的な情報源であるマルコフ情報源の一種です。通常マルコフ情報源では，出力信号の系列が決まれば，どのような状態の遷移を経てその信号が出力されたかが一意に定まりますが，隠れマルコフモデルの場合には，出力された信号からモデル内の状態遷移系列を定めることはできません。逆にこの性質により，同一の隠れマルコフモデルから，多様な信号出力が可能になります。これは，同じ「あ」という音声でも，話者，抑揚，方言，といった様々な要因で，多様な音として発音される性質をモデル化するのに大変都合の良い性質です。

一方，隠れマルコフモデルは，様々な状態遷移パターンを考えることにより，時間信号の時間軸上での伸縮をモデル化することが可能です。図2には，破裂音 /k/ の音声波形と，対応する隠れマルコフモデルの時間推移状態の例を示しています。音素は複数（閉鎖，破裂，気音）の音響的な状態の連続で構成されており，隠れマルコフモデルの状態がそれぞれの部分に対応付けられることが分かります。

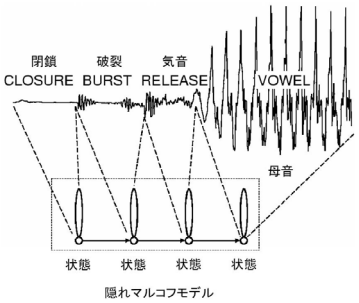

図2　破裂音 /k/ と対応する隠れマルコフモデル

隠れマルコフモデルを利用するもう一つの利点は，音素毎にモデルを学習しておけば，それらを連結することで任意の単語を構成することができることです。日本語の音素はおおよそ40種類と言われていますので，40種類の音素に対応する隠れマルコフモデルを学習しておけば，どのような入力音声にも対応できるわけです。しかし，音素の音としての特徴は前後の音素の影響を受けて大きく変動します。そのため実際の音声認識システムでは，音素毎に隠れマルコフモデルを作成するのではなく，前後の音素との組み合わせを考慮した三つ組み音素（例えば，/kak/ と /pak/ とを区別するなど。）を，単位とすることが一般的です。このような三つ組み音素（トライホン）を利用すると，日本（niqpoN）という単語は，

＃-n+i，n-i+p，i-q+p，i-p+o，p-o+N，o-N+＃

の6つの列に置き換えることができます(*2)。

このように，隠れマルコフモデルを音素単位に作成し，それらを連結した単語のモデルが与えられれば，その単語モデルから認識すべき音声が出力される確率を計算することができます。すなわち，入力された音声がどの単語に近いか，その近さが確率的に与えられることになります。この近さの値を式で表現するならば，ある単語のモデルWが与えられた下で，観測された音声Aが出力される確率ですので，P（A|W）と，条件つき確率で表現できます。

2.4 言語のモデル化

いま音声認識が認識対象とする単語が20単語あったとします。この20単語の中の3単語をつなげてできる文の総数は，20の3乗＝8000文あることになります。しかし，この8000文はどれも同じように出現する訳ではありません。単純に単語をつなげただけでは意味のない文も出現するからです。

例えば，「私」，「は」，「学校」，「に」，「行く」，の5単語について考えると，この5単語の並び方のうち，普

通の日本語として使われるのは，せいぜい

私／は／学校／に／行く

私／は／行く／学校／に

ぐらいで，その他の文はめったに現れないと思われます。つまり，単語の並びには出現しやすい並びと出現しにくい並びが存在します。このような単語の並びの出現確率のモデルは，「言語モデル」と呼ばれます。

現在の音声認識では，特定の3単語が連鎖する確率（トライグラム確率と呼ばれます。）に基づいて，文の出現確率を計算することが一般的です。このモデルを利用することで，任意の単語系列Wの出現確率P(W)を，計算することが可能となります。

三つ組み単語の総数は，語彙の3乗になるため，例えば1万語の語彙を持つシステムでは，10^{12}といった天文学的な数のトライグラム確率を推定する必要があります。このため，トライグラムの推定には新聞記事10年分といった巨大なデータベースが利用されています。

2.5 最適単語列の探索

上記の音声モデル（隠れマルコフモデル）と言語モデル（トライグラム）を用いることで，音声Aが観測された下で，ある単語列Wが出現する条件つき確率P(W|A)を，

$$P(W|A) = P(W, A)/P(A) = P(A|W)P(W)/P(A)$$

と計算することが可能になります。したがって，全てのWに対して上式を計算し，最も高い確率を与えるWを認識結果とすれば良いことになります。しかし，実際には単語列の総数は膨大になるため，適切な方法で計算対象となるWの数を絞りこむ必要があります。この処理は，認識対象として正しそうな系列を「探す」ことで，実現されるため，探索処理とも呼ばれます。

3 音声認識の研究の動向

このような音声認識の基本アルゴリズムを改良する研究が様々な観点から進められています。

(1) 大語彙連続音声認識

認識対象語彙が増えれば，混同しやすい単語の組が増加し，認識誤りが起きやすくなります。同時に，認識処理に必要な計算量が大きくなることも間接的に認識性能に影響を与えます。

この計算量は単語の組み合わせ数に比例しますので，語彙数の増加に伴う計算量の爆発は認識技術の実現にとって致命的です。そこで，認識対象単語の組み合わせ全てについて，確率計算を行うのではなく，様々な処理のレベルで計算を省略する手法が研究されてきました。計算機の高速化の恩恵もあり，現在では数十万単語の語彙を持つシステムが一般的なPC上で，実時間で動作するようになっています。

(2) 環境変動への耐性

音声認識システムの多くは，静粛な環境下や，よく訓練された話者の音声に対しては，非常に高い性能で動作するにもかかわらず，雑音下や，異なる話者に対して性能が大きく劣化することがあります。これは，隠れマルコフモデルの学習に用いられた音声データと，実際の認識対象となる音声の性質が大きく異なるためです。これらの環境の変動に対応するために，様々な研究が進められています。

雑音下で発声された音声から雑音成分を推定して，これを入力音声から減算する方法のために，様々な雑音推定の手法が研究されています。また，音声（話者）と雑音源の場所が離れていることを想定し，複数のマイクロホンを利用して，音声と雑音とを分離する方法についても研究が進められています。さらに，雑音が混入しても，値の変化が少ないような音声の特徴量を探索する試みも多く行われています。

一方，予め作成されている隠れマルコフモデルを，少量の学習データを利用して新しい環境に適応させる方法も重要な研究課題です。この技術を利用すれば，学習時と異なる環境で発声された音声も，高い精度で認識することが可能になります。

(3) 話し言葉・対話処理

新聞などテキストを「読み上げた」音声と，通常の話し言葉とは，使用される語彙や言い回しが異なります。さらに，話し言葉では言いよどみやいい間違いが頻発するため，読み上げ音声で学習された音声モデルで，高い認識性能を得ることが困難です。このような話し言葉に固有な問題を解決するための研究も盛んに行われています。話し言葉には，読み上げ音声に比べて，音響的に多くの変動が含まれていると考えられるため，複数の音声モデルを選択的に併用する方法が研究されています。また，話し言葉のための言語モデルの構築方法も研究が進められていますが，現状では，読み上げ文と話し言葉との間では音声認識の性能はまだ大きく異なります。

一方，音声を利用して対話的なインタフェースを構成する方法の研究も重要な研究課題です。音声認

識を様々な応用システムに適用するためには，認識すべき語彙や言い回しの設計，適切な言語モデルの作成など，多くのサブシステムの作成が必要となります。全体として使い易い音声認識システムを構築するためには，個々のサブシステムの性能が高く，バランスが取れていることが重要です。また，システムの受け答えにどのような文章を利用するか，どのように発話を促すか，といった，ヒューマンファクターに関する研究もこれからの研究課題であり，工学・情報科学だけでなく，認知心理学の分野と連携を取った研究が進められています。

4 音声認識の応用と展望

次に現在音声認識が利用されている主な応用分野と，今後の応用が期待される分野を紹介したいと思います。

(1) 読み上げ文の書き取り

いわゆる口述筆記を，音声認識を利用して行う「音声ワープロ」が，PC上で動作するソフトウェアとして販売されています。利用したことがある方も多いと思います。これらは，数万の語彙を持つシステムで，新聞記事などを利用して予め学習された言語モデルと，大量の話者の音声で学習された音声モデルを内蔵しており，日常ワープロを利用して作成される典型的な文章などであれば，読み上げたとおりに文書を作成してくれます。多くのシステムでは，認識用の音声モデルを利用者の声に合わせるために，音声の特徴を学習する機能が付加されていますが，事前に学習を行わなくても，比較的高い精度で認識を行うことが可能です。

(2) 情報検索

音声認識技術を利用して，音声をテキストに自動変換することができれば，様々な情報処理を音声に適用することができます。例えば会議の様子を録音し，これを文字に書き起こすことで，議事録を作成したり，電話会話の内容を記録に残すことなどが可能となります。一旦テキストに変換された音声に対しては，検索が容易にできますので，放送や通話などに音声やビデオの形式で蓄えられている大量のコンテンツに対して，検索を行うことが可能となります。

(3) ハンズフリー機器操作

例えば自動車の運転中は，視線や手足の操作を運転以外の機器操作に利用することができません。このように，他の作業と並行して機器の操作を行う必要がある場合には，音声を利用することが効果的です。音声認識機能を搭載したカーナビシステムも普及してきましたが，現在では単純な操作コマンドと地名の入力が主な用途となっています。車内のような雑音の大きい環境でも，高い認識性能が得られる技術，運転者に負担をかけない円滑な対話を生成する技術，などが発達すれば，渋滞の回避，レストランの検索や予約といった，高度な情報サービスを，走行中の車内で提供することが可能になります。

(4) ロボットとの対話

ヒューマノイドと呼ばれる，人型ロボットの開発が進んでいます。人型ロボットの開発では，日常生活の場で人間とで共存することを目的に，人間と同様の歩行移動や運搬動作などを行うロボットの実現が目指されています。このようなロボットの操作は，当然人対人同様に，音声言語を利用することが望まれます。このためには，家庭や職場において様々な人の声を聞き分ける技術や，複数の話者が同時に発声した言葉を聞き分ける技術が必要となります。このような，人間の聴覚機能に匹敵するロボットの音声認識機能の研究は，まだ始まったばかりですが，将来が注目される研究分野です。

5 おわりに

音声認識技術は，人間の知性の直接的な表現である「言語を操る能力」を工学的に実現することを目標とする，極めて野心的な分野であり，50年に渡り多くの研究者を惹きつけてきた研究テーマです。さらに現在でも，大学や企業の研究所において，多く研究者により新しい事実やそのモデルが見出されている，活発な研究分野でもあります。大学や大企業だけでなく，日米のベンチャー企業も積極的に研究開発を行ない，新しいサービスを作り出しており，情報科学，工学，言語学，認知科学，など様々な分野の若い学生や研究者が，この分野の研究に興味を持ち，音声認識技術が発展することが期待されます。

注

*1) Cepstrumというのは，SpectrumのSpecの部分を逆から読んだ造語で，周波数領域で表現された信号(spectrum)を，フーリエ逆変換を利用して，時間領域の信号に変換することで得られる信号であることから，このように名づけられた。

*2) #は，単語の境界を表している。また，ここでは，
『(左側の音素)-(中心音素)+(右側の音素)』
で，三つ組み音素を表現している。」