音声認識システムの自律的学習を目指して

工学院 情報通信系
篠崎 隆宏 研究室

篠崎（しのざき） 隆宏（たかひろ） 教授　1976年神奈川県生まれ。東京工業大学大学院情報理工学研究科計算工学専攻博士後期課程修了。米ワシントン大学、京都大学、東京工業大学、千葉大学を経て2013年に東京工業大学総合理工学研究科准教授に着任。2016年より東京工業大学工学院情報通信系准教授。

今や我々に親しい存在となった音声認識であるが、システムの学習にかかるコストが普及の足かせになっている。そのコストを減らすため、篠崎研究室では半教師あり学習という方法を用いてシステムが自律的に学習する仕組みの実現を目指している。ここでは音声認識の仕組みに触れながら、研究室で現在進められている研究について紹介する。

音声認識の抱える課題

皆さんはスマートフォンの音声検索をご存知だろうか。音声検索では、文字をキーボードではなく、私たちの声を使って入力するのだが、「明日の天気は？」「大岡山駅までのルートは？」などの短文であればほとんど正しく認識される。現在、音声検索で使われている音声認識システムは大変精度が良く、実用化されているということもあり技術として成熟されているように思える。しかし今の音声認識システムでは実装まで漕ぎつけていない分野がいくつか存在する。

その一つが、音声認識システムを用いた自然な対話である。音声対話機能を備えたロボット、あるいはSiri[1)]に代表されるようなスマートフォンのAIと会話を行うと、ローカルな話題や新しい単語が正しく認識されないことがよくある。これは人と人との対話であれば多少知らない単語があってもお互いに知識を広めながら対話を進められるのに対し、現在のシステムでは知識を蓄えながら対話を行うことが不可能であるためである。

現在普及している多くの音声認識システムの学習は、教師あり学習と呼ばれる方法に基づいている。この学習法は、音声がどの文字列に対応しているかを逐一教えていく方法である。この方法では、ローカルな情報を学ぶ仕組みが無いために学習の柔軟性に欠け、自然な対話を実現する上で有用でない。

また、教師あり学習は、その学習に大きな手間と費用が掛かるものとなっている。具体的にどの程度の手間があるかをみてみよう。教師あり学習によって高い認識精度を実現するためには、100時間から1000時間、あるいはそれ以上の大量の書き起こしつきの音声データを用いた学習が必要で

1)　Siriは、Apple Inc.の商標です。

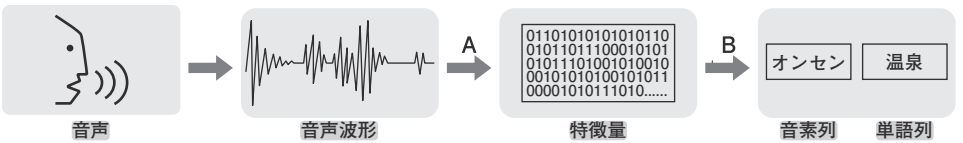

図1　音声認識の基本的な流れ
コンピュータで扱う音素列をここでは理解のためカタカナで表す。

ある。その書き起こしをするためには、音声1時間につき20時間から50時間もの作業が必要で、これは人手で行わなくてはならない。これでは学習を行うのに膨大なコストがかかり、システムの開発後でも定期的にメンテナンスするのに大きな負担となる。

システムの開発費を削減してさまざまなアプリケーションの実現を容易にするためにも、柔軟な音声での対話を実現する上でも、システムがより賢く学ぶ新しい仕組みが必要である。

その新しい仕組みとして今研究が進められているのが、半教師あり学習を利用した音声認識である。半教師あり学習とは、教師あり学習と自律的な学習を組み合わせた方法である。この学習法が有効に作用すると、人手で文字を書き起こす作業が減り、学習の効率が向上することになる。またシステムが新しい語彙を自動で獲得することも可能になる。音声認識システムが今までよりずっと簡単かつ柔軟になるのだ。

ここでは、篠崎研究室がどのように半教師あり学習の研究を進めているかをみていこう。

音声認識の仕組み

先生の研究の本題に入る前に、まず音声認識とは何か、どのように働くことで認識が可能になるかを説明しよう。音声認識の一連の流れは、大きく二つのステップに分かれている（**図1**）。

一つ目は、音声から認識に必要な特徴量と呼ばれるデータを抽出する過程である。私たちの話す声は、コンピュータにとっては数字の連続でしかない。これを次のステップで行う計算に使用するため、音声の周波数や振幅の情報を表すいくつかの実数の集まりを取り出す。この実数の集まりのことを特徴量と呼ぶ。このとき、音声認識に必要な情報だけを集めるように工夫することで、特徴量のデータ量を圧縮することが可能である。データ圧縮を行うことは、認識に不要な情報を取り除き、音声認識の学習の効率を向上させることにつながるのだ（**図1— A**）。

二つ目は、上のステップで出した特徴量を元にそれがどの単語に対応しているか判別する識別の過程である。実際には、特徴量から単語を判別するのではなく、特徴量か単語列を導く前に、音素列の候補を導く判定が必要である。音素とは、特徴量が異なっていても特定の言語の中で人間が同じ音だとみなす音の要素の集合のことである。これらの一連の判別は、特徴量から当てはまる可能性が高い音素列の候補を計算し、音素列から最も当てはまる確率が高い単語を計算するという確率の問題に帰着させることができる（**図1— B**）。

今紹介した音声認識の仕組みは音声認識の枠組みの一つである。そしてこの枠組みでは、一般にAとBの過程が独立しているため、別々に研究をすることができる。先生が行なっている半教師あり学習の研究は主にBの識別の過程である。このBの過程をもう少し詳しくみていこう。

識別の過程は確率の問題だと述べたが、特徴量から音素列を判別するために必要な確率は、音響モデルを用いることで導くことができる。音響モデルとは、各音素に対応して特徴量が生起する確率の分布の集合のことである（**図2**）。このように複数の確率分布の集合したものは確率モデルと呼ばれる（**図3**）。

また音響モデルとは別に、言語モデルという文字列の生起確率を表す分布も同時に用いる。音響モデルが特徴量に関係する確率モデルであるのに対し、言語モデルは特徴量に関係のない確率モデルのため、一見不要なように思えるが、これを利用することで音声認識を一層正確にさせることが

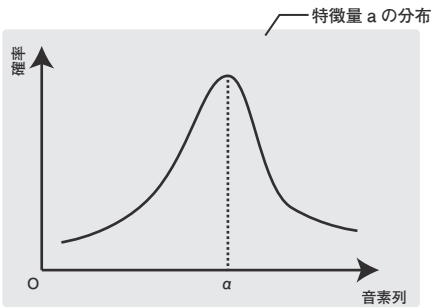

図2　音素に対応する特徴量の確率分布

この分布の場合、音素aに対し最も確率の高いαが特徴量として出力される。

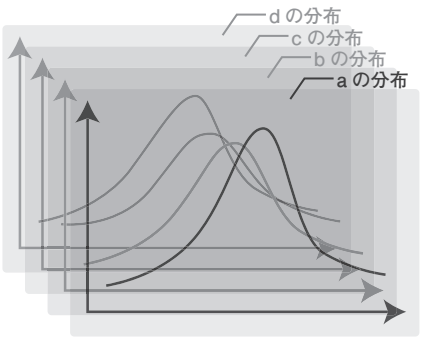

図3　確率モデルの構造

確率モデルは、複数の確率分布の集合である。

できる。例を挙げて考えてみよう。「オンセンニハイル」という音素列を認識する場面で、何らかの原因で認識に失敗し、音響モデルが「オンセン」を「オンセイ」と誤認識してしまったとする。このとき、言語モデルによって「オンセンニハイル」と「オンセイニハイル」という文章が生起する確率を出してみると、明らかに前者の確率が高いため、「オンセイ」という認識は間違いだと判断できる。このように、言語モデルを使うことで、認識の間違いを補正することができるのである。

また、音響モデルにより識別した音素列を「温泉に入る」というような単語列に落とし込むためには、音素と単語についてモデル化する必要がある。単語に対する音素列の対応あるいは出現確率をモデル化したものを発音辞書と呼び、これも音声認識に重要な役割を果たす。これらが代表的な音声認識の一連の流れである。

そして、これらの言語モデル、音響モデル、および発音辞書をデータの自律的な学習により、認識を正確にさせていきたいというのが先生の目標である。

自律的に学習する2つのモデル

さて、ここからは半教師あり学習によって、どのようにしてそれぞれのモデルが確率分布を人手で対応させなくとも学習が可能になるのかをみていこう。

■ 音響モデルの半教師あり学習

具体的な音響モデルの半教師あり学習の流れを説明しよう。まずは、音素列に対する特徴量の確率分布を少量学習している初期音響モデルを作る。その後、新しく与えられた初期音響モデルを用いて、音素列の書き起こしが与えられていない特徴量を認識し音素列に変換する。初期モデルは少量のデータから学習されているため認識精度はあまり良くないものの、新しい特徴量とそれに対応する音素列が得られたことになる。その後、その音素列を書き起こしの代わりに用いて教師あり学習を行い、音響モデルを更新する。この過程を繰り返すことによって音響モデルの精度を徐々に向上させる。これが、従来提案されていた音響モデルの半教師あり学習の流れである（**図4**）。

図4　半教師あり学習の仕組み

しかし、この学習法には、一度間違って認識してしまったものが修正されづらいという欠点がある。先ほどの例をあげると、「オンセン」と認識すべきところを、誤って「オンセイ」と認識する確率が高くなるように分布が作られてしまったとする。すると、更新後の音響モデルにも、「オンセン」の特徴量を「オンセイ」と誤認識するような確率分布ができてしまう。このモデルを用いて、先ほど用いたデータからモデルの更新を再度行うと、そのデータにも「オンセン」の特徴量が出現しているため、誤認識する傾向はますます強くなっていく。このような学習法では、一度生じた誤認識が学習の繰り返しとともに強化されてしまうことになる（**図５**）。

その解決法として音声のデータを2つに分け、2つの音響モデルに交互に学習させる方法が考えられる。この方法であれば、同じ音声が同じモデルに連続して認識されなくなるため、間違いを繰り返すという問題を回避できる。しかし、この方法にもまだ欠点があり、データを2つに分けた分、一度に用いることができるデータ量が半分になるため、データ量に対する学習の効果が減ってしまうのだ。

そこで、先生は学習するデータをなるべく減らすことなく、かつ同じ間違いを繰り返さないようさらに改良した学習法を提案した。次はその手法についてみていこう。

まず、初期音響モデルをk個コピーし、M_1,M_2……,M_kとする。書き起こしのない音声のデータ

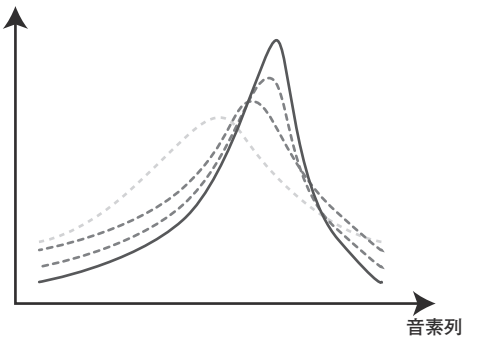

図５　半教師あり学習を行い変化する確率分布
ある傾向ができると、学習するたびにその傾向は強くなっていく。

はk個に分割し、D_1,D_2,……,D_kとする。そして、同じ番号同士で音響モデルを用いて音声を認識させ、認識仮説T_1,T_2……,T_kを作る。認識仮説とは、データと認識結果の対応関係のことである。こうしてできたk個の認識仮説を、それぞれ、自分の番号以外の音響モデルにコピーして学習させモデルを更新する。例えば、1番目の音響モデルは、2～k番目の認識仮説を学習させるといった具合である。そして新しくできた音響モデルを用いて、自分と同じ番号の音声データの認識を行い、できた認識仮説を自分の番号以外の音響モデルに学習させ、モデルを更新する。これらの過程を繰り返して、音響モデルを更新し、精度を上昇させていく（**図６**）。このとき、分割するkの値を大きくとることで、一度の学習に用いる音声データの割合を100%に近づけることが可能になる。この学習を行うことで、モデル推定に用いるデータを

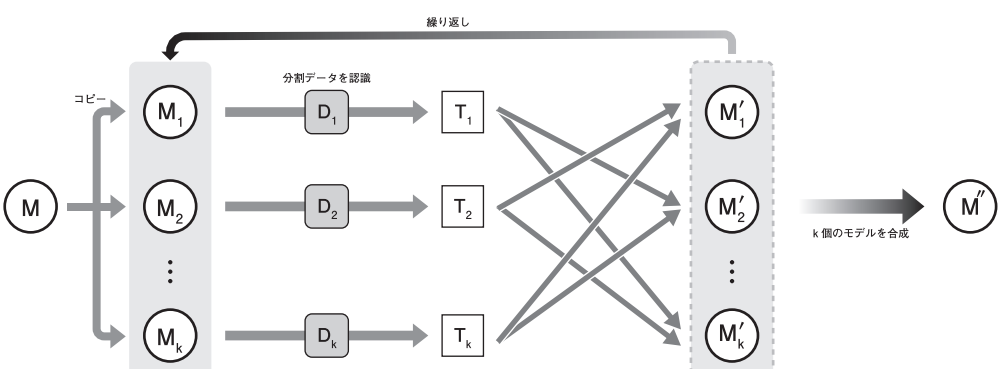

図６　先生が提案した音響モデルの半教師付あり学習

なるべく減らすことなく、かつ連続して同じ特徴量を認識に使用しないようにできる。すなわち同じ間違いを繰り返してしまうリスクを避けることができる。たとえばkの値が10個だとしたら、音声データ全体の90%(=9/10)を、20個なら95%(=19/20)を用いての学習を行うことができる。

以上の仕組みを用いることにより、書き起こしの無い音声データを活用して教師あり学習した音響モデルの性能を従来よりも効果的に向上させることができるようになった。

■ 発音辞書の半教師あり学習

発音辞書も半教師あり学習によって、単語の文字列と音素列の対応を逐一教えることなく、独立した音素と文字のデータを用意するだけで、未知の単語の発音を学習し発音辞書を自動で拡張することが可能になる。この計算にはベイズ推定を用いているので、まずはベイズ推定について説明しよう。

事象Aが起こったという条件の下で事象Bが起こる確率を$P(B|A)$と表すとする。これは、事象Aが起こる条件の下で、事象Aと事象Bが両方起こる確率を示しているので、

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

と表せる。同様に事象Bが起こったという条件の下で事象Aが起こる確率は、

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

と表せる。この2式を合わせると、以下の式が得られる。

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)} = \frac{P(B)P(A|B)}{\Sigma_B P(B)P(A|B)}$$

この式の成立をベイズの定理と呼ぶ。この計算によって、事前に確率が分かっている$P(B)$に対し、観測して得られた情報$P(A|B)$を入れることによって、確率の値を更新し、観測した値に基づく確率$P(B|A)$を求めることが可能になる。

ベイズの定理の意味を具体的な例を出して考えてみよう。例えば、3/4の確率で表、1/4の確率で裏が出るいびつなコインXと、表裏が出る確率がどちらも1/2である正常なコインYがあるとする。この時、ランダムにコインを1枚選び、4回投げたところ(表,裏,表,表)という結果が得られた。この時選んだコインCがXかどうかを推測してみよう。仮にコインを投げた結果が与えられていないとすると、これがXである確率は$P(C=X)$=0.5(=1/2)であるが、今回は結果が与えられているので、より正確な確率を出すことができる。(表,裏,表,表)という結果が出る事象をEと名付ける。求めたい確率は$P(C=X|E)$であり、ベイズの定理を用いると、

$$P(C = X|E) = \frac{P(C = X)P(E|C = X)}{P(E)}$$

のように分解できる。実際に計算すると、この確率を0.628と求めることができる。このように、ベイズの定理を用いて、観測値した値に基づいた確率を求める推定をベイズ推定と呼ぶ。

これを、先生が提案した半教師あり発音辞書学習にあてはめると、音素列Qが与えられたという条件の下で、発音辞書Dの事後確率$P(D|Q)$や、単語列Wの事後確率$P(W|Q)$を求めるという計算に帰着する。これをベイズの定理を用いるとそれぞれ、

$$P(D|Q) = \frac{\Sigma_W \ P(W)P(D)P(Q|W,D)}{\Sigma_{D,W} \ P(W)P(D)P(Q|W,D)}$$

$$P(W|Q) = \Sigma_{Q,D} P(W)P(D)P(Q|W,D)$$

となる。この式を評価するためには、言語モデル$P(W)$に加え発音辞書の確率モデル$P(D)$を設計するとともに、効率的な計算アルゴリズムを実現する必要がある。このために先生はノンパラメトリックベイズ法と呼ばれる手法を応用し、複数の無限分布を組み合わせた確率モデルを考案した。またギブスサンプリングと呼ばれる手法を用いた計算アルゴリズムを提案し、直接関数の形で求めるのが難しい確率分布を生成したサンプルの集合によって表現するようにした。このようなプログラムを実装することによって、実際にベイズ法に基づいた半教師あり辞書学習が可能であることを世

界ではじめて実証した。

このように、半教師あり学習を用いることで、2つの情報の対応関係を直接与えなくても、対となる情報を一部与えれば、あとは自律的に2つの関係を学習することができる。先生は、音声認識においての半教師あり学習が上で挙げた音響モデルと発音辞書においてうまくいくことを実証した。将来、精度の高い音声認識が、今まで以上にさまざまな場面で使われるかもしれない。

柔軟な学習がもたらす未来

今まで学習といえば教師あり学習が主流であったが、半教師あり学習を用いることで学習のコストを下げるだけではなくより柔軟な学習をすることが可能となる。

半教師あり学習は何も音声と文字だけではなく音声の代わりに画像を使い、画像認識と組み合わせて文字を学ばせるという場面にも用いることができる。また、コミュニケーションを軸にした学習法もある。これは、ある指示を与え、それが正しく実行できるかどうかを評価することで意味理解がどこまで可能になっているか判別する学習法である。

このようなさまざまな手掛かりを利用した多角的な学習は、さながら人間が言語を学んでいく過程をみているようである。我々が母国語を習得する際は、何も単語帳を使って言葉を逐一学んできたわけではない。我々は絵本やテレビなどの視覚的な情報や耳にした会話の膨大な蓄積などを手掛かりに、言葉を学んできた。半教師あり学習から柔軟性のある学習法を研究することは、我々がどのように言葉を学ぶのかということを研究することにつながる。つまり、性能の高くコストが安いシステムを作るという工学的な観点からも、人間の柔軟な学習法をコンピュータ上で再現するサイエンティフィックな観点からしてみても、半教師あり学習は大変興味深い分野だと先生は語る。

さらに、人手に全く頼らない学習法として教師なし学習というものも存在する。この学習では対となる情報を一切与えず、2つの事象の関係を推測する。これが実現できれば、システム設計者が想像もしなかったような関係を探すことも可能になる。もちろん、これは半教師あり学習以上に実装が難しいところではあるが、これにも意欲的に取り組んでいきたいという。

先生が音声認識の研究を始めたきっかけは、高校時代までさかのぼる。先生はもともと電子工作が好きで、古いテレビなどから部品を集め簡単な金属探知機や高電圧発生装置などを作っていた。そしてさまざまなハードウェアを作っていくうちにもっと面白いものが作りたいと思うようになり、当時ではまだ怪しい分野とささやかれていた人工知能に興味を持ち始めた。そして人工知能研究の入り口として音声認識を選んだのだという。

先生の究極的な目標は、さまざまな人間の学習法を組み込んだ人工知能を積んだロボットやエージェントを作ることだ。もしこの仕組みが実現すれば、世界中の人々との膨大なコミュニケーションを用いて人工知能が自然に言葉を覚え、対話を繰り返しながらまるで本物の人間のような学習をすることができる。

人間のような学習を実現するにはまだまだ課題が多い分野ではある。実際、より人間に近いとされる教師なし学習については国際的にもやっと地に着いた研究が始まったといえる状態で、人間のような学習ができるというのは今はまだ夢物語の段階だと先生は語る。そのため、先生はこれまでみてきたように目標までの道のりの中で着実に進められそうな分野から研究を進めたり、あるいは限定的な条件の中でもシステムが実現できるようにするため、試行錯誤を繰り返したりして一歩ずつ目標へと近づこうとしている。

執筆者より

快く取材をひき受けてくださった篠崎先生に心よりお礼申し上げます。取材時には難しい研究内容を、分かりやすく簡潔に、それでいて東工大生に読み応えのあるように掘り下げて教えて下さいました。また、執筆の間も記事に的確なアドバイスをしていただき、そして、私からの数多くの質問にも丁寧に返答していただき、大変感謝しております。　（堀本 遊）