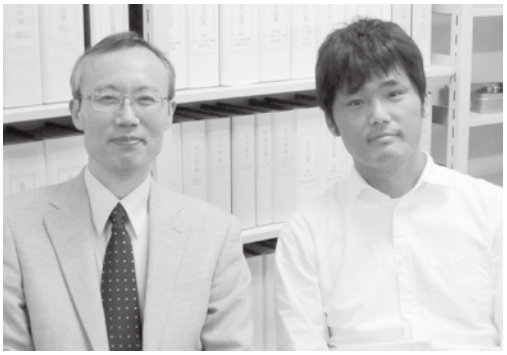
In Laboratory Now

研究室訪問 4

一歩先の音声合成を目指して

小林 隆夫 研究室～物理情報システム専攻



小林 隆夫 教授　　能勢 隆 助教

音声合成の技術は、我々の身近なところではバスや鉄道の車内放送、カーナビゲーション、天気概況、株式市況などで用いられている。現在実用化されている音声合成では、生成された音声の品質が悪いか、あるいは品質が良くても特定の人の声しか生成できず、また感情のこもっていない声しか生成できなかった。

小林研究室では、隠れマルコフモデルを用いて、色々な人の音声や感情がこもった音声を合成するための研究を行っている。本稿では、先生らが行っている三つの研究について紹介する。

テキスト音声合成とは

小林研究室では、コンピュータと人間の間で音声を用いて情報をやり取りするための処理について研究しており、その中でも音声合成についての研究に力を入れている。音声合成とは、コンピュータを用いて人間の音声を作り出すことである。音声合成にはいくつかの様式があるが、小林先生が研究しているのは、テキストを音声へ変換するテキスト音声合成と呼ばれるものである。

従来の一般的なテキスト音声合成は、テキスト解析、韻律生成、合成単位選択、波形生成の四つの段階に分けることができる。最初の段階のテキスト解析では、与えられたテキストの読み、間、アクセントなどの特徴を、言語辞書というデータベースを参照して決定する。

次に、テキスト解析で得られた情報を基にして韻律生成と合成単位選択が行われる。韻律生成の段階で、声の高さ、大きさ、イントネーションやリズムといった、韻律と呼ばれる特徴を生み出す。韻律の調整によって、ある程度話者の感情を表現することができる。ここで重要となるのが、人間の声の高さは一定になっておらず、時々刻々と変化しているということである。そのため、人間の音声における韻律の変化を再現することで、より人間らしい音声を合成することが出来る。また韻律生成をするのと同時に、繋ぎ合わせることで音声を形成するような短く区切られた波形（合成単位）を、合成単位選択で選び出す。

最後に、選び出した合成単位を韻律の情報に従って繋ぎ合せる。こういった過程を経て、与えられたテキストを音声として出力することができるのだ。

テキスト音声合成における研究目標は、自然で表現豊かな音声を合成することである。しかし、色々な個性を持った声の合成、様々な感情の表現、TPO による喋り方の違いの表現を実現するのは難しい。小林研究室では、これらを実現できるように、従来のテキスト音声合成の方法を改良するための研究を行っている。小林研究室で行われている研究のうち、本稿では、平均声と呼ばれる、複数の話者の平均的な特徴を持った音声に基づく音声合成に関する研究、多様な感情を表現する音声合成に関する研究、声質変換と呼ばれる、任意の話者の声を他の話者の声に変換することに関する研究を紹介する。

新しい音声合成、HMM 音声合成

テキスト音声合成の分野において、20 年ほど前から研究が始まり、現在の主流となっているのがコーパスベースの音声合成というものである。この方法は、事前に一人の人間の音声を大量に録り溜めておき、その大量の音声波形を使って音声合成を行うというものである。音声を合成する際には、目標とする韻律に適して、かつ滑らかに接続できる波形をデータの中から探し、それらを組み合わせる。この方法が登場したことによって、より自然な音声を合成することができるようになった。

しかし、この方法には、品質の高い音声を合成しようとすると、非常に大量のデータが必要になるという欠点があった。その上、どれだけ大量のデータを集めたとしても、繋ぎ目が不自然になる部分が出来てしまい、あらゆる場合の音声を表現できるわけではなかった。

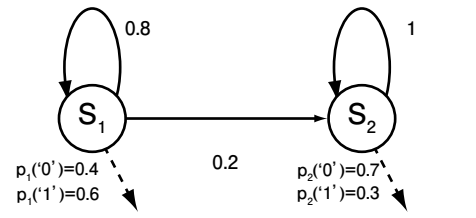
そこで最近多く使われるようになったのが、コーパスベースの音声合成を隠れマルコフモデル（Hidden Markov Model ; HMM）で改良した、HMM 音声合成という方法だ。HMM とは、時間的に性質が変化する信号のモデル化に使われる確率モデルである。これを時々刻々と変化する人間の声の波形に対して応用したのが、HMM 音声合成である。小林研究室の研究を紹介する前に、その基盤となる HMM 音声合成について紹介する。

音声を扱う際に実際に使われる HMM について説明する前に、一般的な HMM について説明しよう。HMM は状態と確率的に起こるその遷移を扱うモデルである。HMM の特徴として、次に起こる遷移の確率が過去と現在の状態により決定されることと、実際にどのような遷移が起こったかがわからないことが挙げられる。音声を扱う際に主に用いられるのは一次マルコフモデルという、次に起こる遷移の確率が現在の状態のみにより決定されるものである。実際に起こった遷移の過程はわからないが、遷移に伴って出力されるデータ（シンボル）は観測できる。シンボルとは、ある状態に遷移するごとに、各々の状態ごとに決定される確率に従って出力されるものである。シンボルの集まりを観測することで、どのような経路をたどった確率が一番高いのかを計算し、実際に行われた遷移の過程を推定することができる。

音声合成や音声認識の分野で一般的によく用いられる HMM に、left-to-right モデルというものがある。これは、状態の遷移が一定方向に進む一次マルコフモデルである。HMM 音声合成でも left-to-right モデルを使っている。

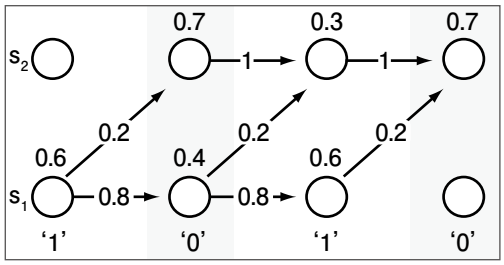
ここで簡単な left-to-right モデルを例として、たどった経路の推定をしてみよう（**図 1**）。S_1、S_2 という二つの状態があり、初めは S_1 の状態だとする。単位時間後にシンボルを出力した後も S_1 の状態のままである確率を 0.8、S_1 から S_2 に移る確率を 0.2 とする。S_2 に移った後は S_1 に戻ることはない。出力するシンボルを 0 と 1 の二つとして、S_1 では 0.4 の確率で 0 を、0.6 の確率で 1 を出し、S_2 では 0.7 の確率で 0 を、0.3 の確率で 1 を出すものとする。

このモデルにおいては、初めにシンボルを出したときは S_1 の状態、最後にシンボルを出したときは S_2 の状態となっており、このときシンボルの時系列 '1010' が得られたとする。時系列とは、時間経過に従って計測されるデータ列のことである。ここで、S_1 から S_2 に移ったのは一番目、二番目、最後の遷移のいずれかである（**図 2**）。一番目に S_1 から S_2 に移ったとすると、S_1 で 1 を出してから S_2 に移り、そのあと S_2 で 0、1、0 の順にシンボルを出したことになる。このようなことが起こる確率は、出力の確率と状態の遷移確



状態遷移を図を用いて表す。
ここで、S_1、S_2は状態を、矢印（実線）は状態の遷移を表す。
この図では、S_1の状態でシンボルを出力した後、0.8の確率でS_1の状態にとどまり、0.2の確率でS_2に遷移することを示している。
また、矢印（破線）は出力を表す。「p_1('0')=0.4」は、状態S_1の時、シンボル0が0.4の確率で出力されることを意味する。

図 1　left-to-right モデル



状態の遷移が図1で表されるモデルにおいて、'1010'というデータ列が出力された時、考えられる状態の遷移は、このようになる（矢印上の数字は状態遷移の確率、状態を表す円の上の数字は各シンボルを出力する確率を表す）。

図2　left-to-right モデル

率を考慮すると、順番に積をとって 0.6 × 0.2 × 0.7 × 1.0 × 0.3 × 1.0 × 0.7=0.01764 となる。同様の計算で、S_1 から S_2 に切り替わったタイミングが二番目のときは 0.6 × 0.8 × 0.4 × 0.2 × 0.3 × 1.0 × 0.7=0.008064、最後のときは 0.6 × 0.8 × 0.4 × 0.8 × 0.6 × 0.2 × 0.7=0.0129024 となる。以上により、「1010」というシンボルの時系列が出力される確率は、それぞれの和を取ることで 0.0386064 だとわかる。また、切り替わりのタイミングは一番目である確率が最も高いということもわかる。

音声合成で HMM がどのように適用されるか説明しよう。HMM 音声合成では、コーパスベース音声合成で合成単位をモデル化する際と、合成単位から韻律及び読みやアクセントの情報を生成する際に HMM を用いる。上記の例ではシンボルは１と０といった孤立した値だったが、連続値で表された音声の特徴量ベクトルそのものをシンボルとして取扱う連続 HMM が使われる。特徴量ベクトルは、周波数ごとの音の強さや声の高さといった音声の特徴を表す。音声合成において使われる HMM は、１音素を３ないし５状態に分けた left-to-right モデルである。音素とは、聴覚的に区別できる音声の最小単位で、母音や子音のことである。当該音素が同じでも、前後の音素によって特徴が変わってくるので、前後の音素も考慮する必要がある。そのため音声認識や音声合成では、一つの同じ音素でも前後の音素が異なるものは別のものとするトライフォンという単位がよく用いられる。各状態は、音素の中で似たような特徴を持った部分の集まりを表している。

HMM 音声合成では、テキスト解析より後の段階が従来の音声合成とは異なる。従来別々に扱っていた読みの情報と韻律の情報が，合成単位 HMM 内に共にモデル化されているのである。HMM 音声合成では、テキスト解析の後に、読みやアクセントの情報を記述したコンテキストラベル列というものを一旦作成し、それを基に合成単位 HMM を選択する。合成単位 HMM は、あらかじめ録り溜めた音声を基にトライフォンを用いて作成される。その合成単位 HMM を繋ぎ合せて文章 HMM を作る。文章 HMM からイントネーション、リズム、声の大きさ、周波数ごとの音の強さ等のパラメータを生成する。このパラメータを元にして、人間の発声の構造を近似するモデルに基づいて、波形を生成する。

音声合成に HMM を用いることで、波形接続に基づく従来のコーパスベース音声合成に比べ、より少量の音声から自然な韻律を持ち滑らかに聞こえる音声を合成できるようになるというメリットがある。

多様な音声合成

HMM 音声合成の応用

ここからは、小林研究室が行っている、HMM 音声合成を応用した音声合成に関する研究について紹介する。

従来のコーパスベースの音声合成には、いくつかの課題があった。第一に、大量の音声データが必要だったという課題があげられる。この方法では、任意の話者の発する音声を合成しようとする度に、その人の膨大な量の音声を新たに録音する必要があり、非常に手間がかかってしまう。少なくとも数時間程度の音声データを用意する必要があり、製品レベルの品質にしようとすると数十時間程度の音声データを用意する必要があるのだ。これでは、例えば有名人の音声を合成したいという場合に、数十時間もの録音を依頼することになってしまう。そこで小林研究室が提案するのが、平均声に基づく音声合成というものである。

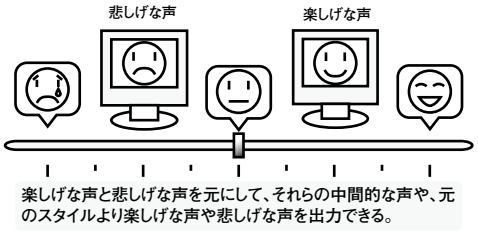


図３　比率によるスタイルのコントロール

平均声とは、複数の話者の音声データから作られる音声で、複数の話者の平均的な特徴を持っている。平均声に基づく音声合成は、HMM 音声合成が基盤となっている。通常の HMM 音声合成では、目的の話者の声のみを、例えば 30 分ほど録音して、それを基に音声を合成する。それに対して、平均声に基づく音声合成ではあらかじめ、一人の声ではなく複数の男女の声を例えば各話者15分ないし30分収録しておく。これらの音声データを基にして、平均声モデルと呼ばれる音素モデルを作り出す。そして、目標とする話者の一分に満たない音声データを録り、それを用いて平均声モデルを目標の話者の音声に近づけるように適応させることで、十分に目標の話者の特徴をとらえた音声を合成することができる。これにより、録音する必要がある目標の話者の音声の量を、従来のコーパスベースの音声合成に比べ格段に減少させることに成功した。

第二に、従来のコーパスベースの音声合成では、自在に感情を表現することができないという課題があった。従来の方法でも、テキスト音声合成で生成される音声の品質は向上しており、人間が話したような声が作れるようになってきた。しかし、それにより合成された音声は、豊かな感情表現を持たないアナウンサーのような声だった。感情がこもった声や話し方が特徴的な声は、普通の声よりも変化が激しいものとなる。そのため、そのような話し方をした音声データをさらに大量に用意する必要があるが、あらゆる話し方の音声を全て用意することは実際には不可能であり、不連続な繋ぎ目を持つ部分が出来てしまう。それに対して、HMM 音声合成を応用すれば、少量の音声を用意するだけで、感情がこもった声を合成することができるようになった。

小林研究室では、感情や話し方をスタイルと呼んで、様々なスタイルの音声を合成することについて研究している。あるスタイルを持った声を元に HMM で学習させることにより、そのスタイルを持った音声を合成する。HMM における学習とは、それぞれの状態で、遷移する際の確率や、シンボルの発生する際の確率を求めることを意味する。**図３**のように、楽しげな声と悲しげな声の間を取った声や、通常よりもさらに楽しげな声を出力することもできる。この方法には、実際には用意していないスタイルを持った音声を出力することができるというメリットがある。

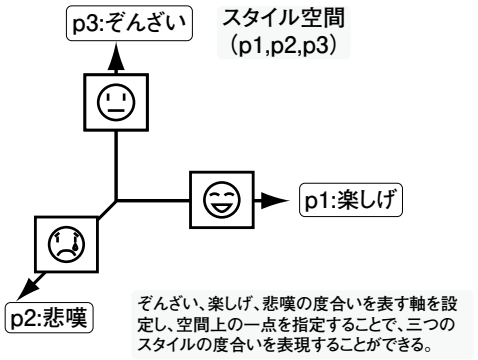


図４　スタイル制御

小林研究室ではこの考え方を発展させ、スタイル空間と呼ばれる多次元空間で複数のスタイルの度合いを表現する、スタイル制御という方法を考案した。例えば**図４**のように、ぞんざい、楽しげ、悲嘆の度合いを表す軸を設定し、空間上の一点を指定することで、三つのスタイルの度合いを表現することができるというものである。

スタイル制御では、話している間にだんだん感情が変わっていくような音声も表現することができる。スタイル空間上で、音声のスタイルを表す点を動かしていくことによって、時間とともにスタイルが変わる音声を合成できるのだ。例えば、初めはぞんざいに、その次は悲しげに、最後は楽しげになるような、従来の方法では表現できなかった音声を出力することが可能になる。

音声合成以外にも、スタイル空間を利用したスタイル認識というのもある。これは、誰かが発した音声のスタイルをスタイル空間上に当てはめた場合に、それがどこに位置するかを推定するというものである。何人かの人に実際に音声を聞いてもらう評価実験を行って、この推定はある程度確かであるということがわかっている。

HMM に基づく声質変換

小林研究室では、テキストを元に音声を合成するのではなく、ある人の声を他の人の声にするという声質変換についても研究している。もともと声質変換は、色々な人の声を出力するために提案された技術である。カラオケなどで使われるボイスチェンジャーのようにリアルタイムで変換することに力を入れた方法もあるが、小林研究室では単なるフィルタとしてのものではなく、声質をより目標とする人に似せることに力を入れている。

ここで、声質変換について説明しよう。音声の音色はのどや口で形成されるため、のどや口の形は音声の個性を出すのに関わっている。声質変換の仕組みは、ある人の声と他の人の声との間でのどや口の形の特徴量や声の高さを表す特徴量などを変換するというものである。小林研究室では、声質変換の際に HMM を用いる方法を提案している。小林研究室が音声合成で培ってきた HMM の方法を声質変換に利用するのだ。HMM を用いた声質変換では、音声から話している文章と韻律を抽出する際に HMM を使い、HMM 音声合成でそれらの情報を元に音声を合成する。

声質変換に HMM を用いることで、従来の方法の声質変換がもっていた三つの問題点を解決することができる。

一つ目の問題点は、話者依存性があるということである。従来の方法では、声の特徴が似ている人同士だとうまくいくが、性別が違う人同士のように声の特徴が大きく違う人同士で変換すると、どうしても品質が悪くなってしまう。従来の方法では音素より細かく分割してモデル化していたために、音声の特徴量の時間的な変化を正しく表現できなかった。HMM を用いることで特徴量の時間的な変化を表現でき、話者の声の個性を正確に捉えられ、話者の変化に応える声質変換をすることができるのだ。

二つ目の問題点は、必要な音声データの量が多くなってしまうということである。高品質な声質変換を行おうとすると、膨大なデータ量が必要となってしまう。これについては、従来の声質変換と HMM を用いた声質変換とを比較する評価実験で、後者の方が少ない音声データで良い品質の音声を出せることがわかっている。

三つ目の問題点は、元の話者と目標の話者について、同じ文章を用意しないと声質変換をすることができないということである。例えば、元の話者と目標の話者について一時間の音声データをそれぞれ用意しても、二つの音声データが同じ文章でなければ変換することができなかった。一方、HMM を用いると音素ごとにモデル化することができるので、異なる文章を用意してもモデルの学習を行なうことができる。このように、HMM に基づく声質変換は、従来の方法よりも優れた声質変換を実現するのだ。

音声合成や声質変換についてのこれからの課題としては、品質を向上させることが挙げられる。ある程度人間らしい音声を作り出すことはできるのだが、日常会話のような音声を合成して人間並みと呼べるレベルまで近付けるのは難しい。人間が文章を読み上げるときと、考えながら話したり砕けた感じで話したりするときとでは、声の抑揚などが全く異なる。ところどころに間があいたり、言い直しや言いよどみが含まれたり、そもそも話している内容が正しい日本語の文章となっていないことすらある。これらは人間らしい会話を表現する上で重要なのだが、これらの要素を組み込もうとしても、合成された音声はなかなか人間らしくならないのだ。音声合成や声質変換についての課題が、小林研究室の研究により解決されることを期待したい。

本稿の執筆にあたり、小林先生には音声合成などについて、非常にわかりやすく説明していただきました。中には難解なお話もありましたが、どれも興味深い内容ばかりで、知的好奇心を刺激されました。

末筆になりますが、ご多忙な中、幾度にもわたる研究室訪問や質問などに快く応じてくださり、ありがとうございました。小林研究室の今後のご活躍を心よりお祈り申し上げます。

（坪井　祥紀）