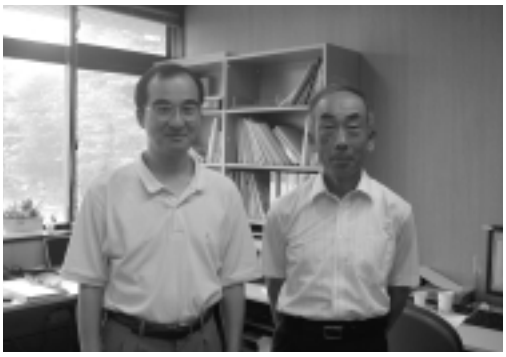
In Laboratory Now

研究室訪問３

音・動画、その意味とは何か？

古井・篠田 研究室～計算工学専攻



篠田 浩一 助教授　　古井 貞熙 教授

コンピュータが普及し、新しい情報が次々と生まれてくる現代、膨大な情報の中から、誰もが必要な情報に容易にアクセスできる時代を築き上げていかなくてはならない。そのためには、個々の情報のもつ意味をコンピュータが自動で認識し、分類しておく必要がある。音、動画、文字、情報は多種多様であるが、それらの真の意味を如何に認識するかが重要な課題である。

今回訪れた古井・篠田研究室では音声に関する研究を中心に、動画像認識への応用まで様々な研究が成されている。

音声認識の今　～音の特徴を捉える～

音、それは日常生活において最も身近で重要な情報源といっても過言ではない。人間は意味をもたせた音を発声することで他人に情報を伝達することができる。しかし、コンピュータが音の持つ意味を的確に認識することはまだできていない。最近では、カーナビゲーションや音声での文章入力などに音声認識システムがみられるようになってきたが、これらは限られた状況下においてのものに過ぎず、未だ多くの研究課題が残っている。例えば、講義やインタビュー、会話などにおける「話し言葉」を認識することや様々な言語に対する音声認識は非常に難しい。SF映画によくあるロボットとの音声対話や多言語の音声自動通訳は音声研究の夢である。

今回訪れた古井・篠田研究室では、そのような夢の実現を目指して、世界15カ国もの国々から学生が集まり、様々な言語に対して音声に関する研究を行っている。日本語や英語はもちろん、中国語やインドネシア語、さらには世界的にも例をみないアイスランド語までもが研究の対象となっているのである。ここでは、音声に関する研究の中から音声認識について紹介しよう。

音声認識の対象となる言語は様々であるが、その核となる概念は一つである。実はどの言語においても音の最小単位が音素であるという共通点が音声認識の鍵となる。音素とは日本語においては母音や子音のことで、例えば「音」という単語は/o/,/t/,/o/の三つの音素から成っていると言える。そこで、音声認識をするために、まず音声を分析して音素の特徴を捉えていく。

図１は男性と女性がそれぞれ/a/と発音した時のスペクトログラム（声紋）である。横軸が時間軸、縦軸が周波数軸で、周波数成分の時間変化を濃淡で表わしている。声には個人差があり、二つのスペクトログラムが完全に一致することはないが、よく見比べると男女共に約1.5kHz付近の周波数成分が強くなり、似たような形をしていることが見て取れる。そこで、これを手掛かりにして、一般的な/a/の形を捉えていく。

実際にコンピュータが/a/を認識するためには、まず/a/のスペクトログラムの形を数値的に表さなくてはならない。その概略を説明しよう。まず、**図２**のように、連続的な音声波をアナログ/デジタル変換によって離散的な数値に変換する。次

に、音声波をある短い時間（20～30msec 程度）で切り出し、その間の周波数成分を求め、その強さを特徴ベクトルとする。さらに、切り出す時間を少しずらし、特徴ベクトルの時系列（時間的な並び）を得る。このようにして作られた特徴ベクトルの時系列はスペクトログラムの形によって決まるので、これによって形を捉えることができるのである。

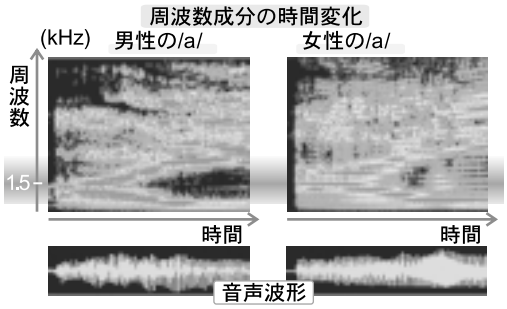
現在の音声認識は、このような音素の形をコンピュータがあらかじめ学習しておき、それを基に認識を行うという手法である。学習は、様々な人の音声データとそれを書き起こしたテキストデータ（合わせてコーパスという）を基にして行う。また、人によって異なるものを扱うには確率統計的な手法が最も適している。そこで、学習には隠れマルコフモデル（HMM）という確率モデルを用いる。HMMは特徴ベクトルの移り変わりを確率的に表現するためのモデルで、各音素がどのような形であるかを学習しておくのである。

しかし、単独の音素を学習するだけでは、音素が連続的につながった単語や文をうまく認識することはできない。なぜなら、音素間の“つながり”の部分が考慮されていないためである。そこで、音素の前後関係も考慮したトライフォンという音素単位を用いる。例えば、/oto/と/ato/の/t/は別ものだと考えるのである。前の音を左に、後の音を右に書けば、

/oto/ = #ot + oto + to#

/ato/ = #at + ato + to#（但し、#は無音）

のように単語をトライフォンの系列で表すことができる。この場合、前後の音素を考慮するため、音素の種類が数十程度であってもHMMを数千種類用意しなくてはならないが、単語を単純にトライフォンの系列として表せるので、何万もの単語ごとにHMMを用意する手法に比べると、効率が良くなるのである。



声紋は一致する事はないが、似通う部分がある

図１　/a/のスペクトログラム（声紋）

学習が済めば、いよいよ認識が可能となる。認識の際は入力音声から特徴ベクトルの時系列をつくり、各HMMがその系列を生成する確率を求め、さらに文法的な要素なども加味して、最も確率の高いものを認識結果とする。ここまでが音声認識の大まかな流れである。

これらの手法を用いることで、音声データを認識し、文として出力できることはもちろん、学習に用いるコーパスが十分な大きさであれば、頑健な認識ができるようになる。実際、ニュースなどにおける原稿の読み上げ音声を認識すると、その認識率は90%以上になるという。

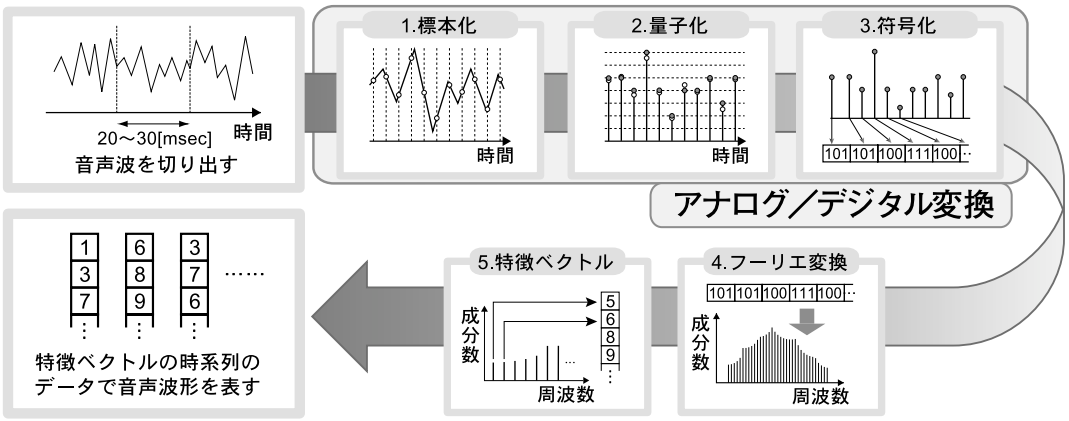


図２　音声の特徴ベクトル

話し言葉音声認識

音声認識において、現在大きなテーマとなっているのは話し言葉の音声認識である。講義やインタビュー、会話での話し言葉を認識することは非常に難しく、先程90%以上であった認識率も約40%にまで低下してしまうほどである。

古井・篠田研究室では話し言葉音声認識の実現に向けて、「話し言葉工学」プロジェクトとして国立国語研究所や通信総合研究所と合同で五年間研究を行った。そこでは1400人もの人々が自由に話した合計650時間の音声データとその内容を書き起こしたテキストデータを集めて、話し言葉コーパスを構築した。このコーパスは質・量ともに世界最大レベルのものであり、世界的な注目も浴びている。そして、現在もそのコーパスを利用して、話し言葉音声の分析や認識などに関する研究が続いている。

話し言葉音声の分析では、話し言葉と書き言葉（読み上げ音声）の違いについて研究を行っている。話し言葉には書き言葉にはない特徴がいくつか存在する。特徴のひとつは「話し言葉の音声は原稿を読み上げた音声に比べて、音そのものがあいまいになってしまう」ということである。同じ/a/という音を発音しても話し言葉では/a/と他の音/e/などが似通ってしまうというような現象である。研究室ではこの「あいまい」という捉えづらい違いを数値的に分析することに成功した。このことは、各音素の波形を多次元のベクトルで表したものを、空間上の点として捉えると違いが分かりやすく説明できる。**図3**は音素空間のイメージ図で、実際には多次元の空間を2次元の平面に射影したものである。例えば、**図3左**の/a/の部分が音素/a/の領域で、男性の/a/も女性の/a/もこの領域の中に含まれている。同じ/a/でも人によって違いがあることを、領域の広がりで表わしているのである。

書き言葉と話し言葉では領域の広がり方が異なるのである。**図3中央**のように話し言葉ではどの音素も原点よりに広がりが増して、音素と原点との距離が縮小してしまう。この距離の縮小率を求めたものが**図3右**である。縮小率はその音素がどの程度あいまいになったかを数値で表わしているといえる。このような話し言葉音声のコーパスと分析を基に、音声認識を行った結果、認識率は約80%にまで向上したという。

今回は話し言葉音声の分析について紹介したが、古井・篠田研究室ではHMMのようなモデルそのもの改良や音声認識の結果を自動的に要約する研究も行われている。また、話し言葉音声認識の更なる課題は未知語への対応やアクセント・イントネーションの利用などといったことである。講義の内容などを認識し文字化できるようになれば、今まで困難であった音声データの検索も容易になる。さらに、話し言葉のスムーズな認識は対話ロボットや通訳にも応用していくことができるのである。これからますます音声情報が注目を浴びることであろう。

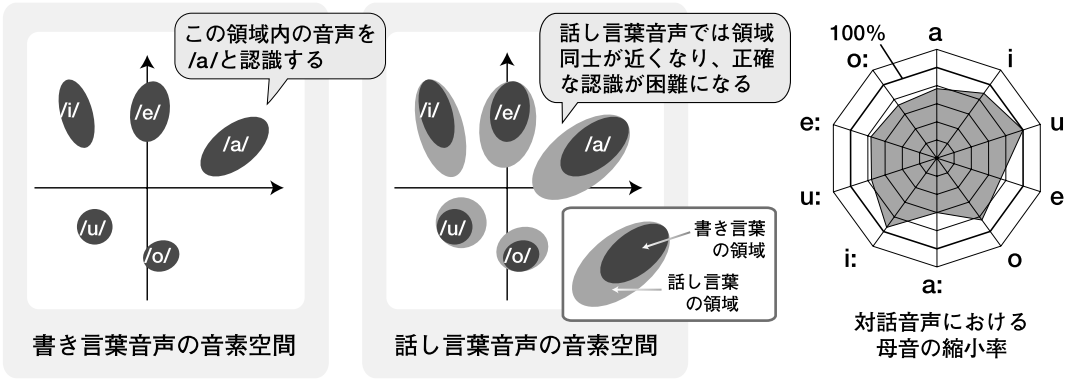


図3　書き言葉と話し言葉の音素空間

動画像認識への応用

「DVDに野球中継を録画したはいいが、全部見ている時間がない。あの選手のホームランシーンだけでも検索して見ることができれば・・・」

動画の内容を検索できれば便利であるが、検索のためには動画の内容を認識しなくてはならない。一章・二章では音声研究に関して紹介をしてきたが、古井・篠田研究室ではまた新たな観点から“認識”についての研究を行っている。そのひとつである「動画の自動インデキシング」という研究について紹介しよう。

インデキシングとは索引付けの意味であり、現在研究の対象としている野球動画でいうと、ホームランや三振といったシーンを自動的に認識し索引付けることが研究の目的となる。インデキシングを行えば、動画検索も容易になり、テレビ局の動画編集などにも利用することができる。従来はこういったインデキシングを人が手動で行っていたのであるが、それには非常に多くの時間とコストがかかり、決して効率的とは言えなかった。そこで、インデキシングの自動化の研究が進められているのである。

まず、研究対象である野球動画の構成をみておこう。野球動画の構成は**図4**ようで、1コマごとの静止画であるフレームが連なってできている。また、一つのカメラで撮影されたフレームの集まりをショット、さらにショットの集まりをシーン、シーンの集まりである野球動画全体をゲームと呼ぶ。例えば、**図4**はあるゲームにおけるホームランのシーンの構成を示している。このホームランのシーンはピッチャーがボールを投げてバッターが打つまでの投球&打撃ショット、打球をカメラが追うショット、その時のピッチャーの憂鬱な表情を撮ったショット、最後に打者がホームインをするショットという四つのショットで構成されている。個々のショットを見ただけでは、何のシーンかは分からないが、いくつかのショットが時間的に並ぶことによって意味のあるシーンを構成するのである。このように、動画においてはシーンがショットの時系列となっているので、それを手掛かりに認識を行っていく。

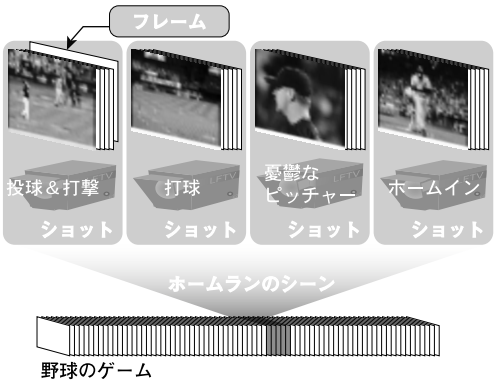


図4　野球動画の構造

古井・篠田研究室では、世界でも例をみない、新手法の動画像認識の研究が行われている。それは、音声認識の技術を応用した動画像認識である。音声と動画、その接点は「動き」にある。どちらも時間的な「動き」に大きな意味があるという共通点があり、構造が類似しているのである。ここで、音声認識の技術を応用するために、音声と動画の対応関係を考えると、**図5**のように音声認識における音素が動画像認識におけるショット、単語がシーン、文がゲームに対応していることが分かる。この関係利用することで、音声認識と同様の流れで野球動画のシーンを認識していくことができるのである。

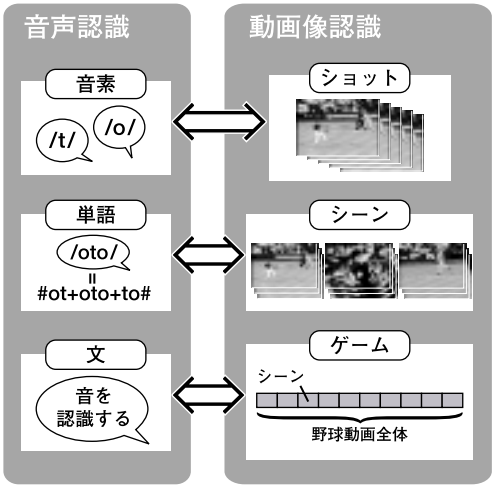


図5　音声と動画の対応関係

具体的な手法を紹介しよう。まず音声データと同じように、動画像データから特徴ベクトルを抽出する。特徴ベクトルは**図６**のように三種類の方法を用いて抽出している。**図６左上**では、動画から低周波成分を取り出し、その成分30個を特徴ベクトルの成分に割り当てている（つまり、30次元の特徴量である）。**図６左下**では2枚の連続した画像の差分の低周波成分から30次元の特徴量を得ている。**図６右上**では人や物の動きとズームの度合いから５次元の特徴量を得ている。

次に学習のフェーズではあらかじめラベル付けされた動画を基に音声と同じくHMMを用いる。これによって、ホームランとはどのようなシーンなのかをコンピュータが学習するのである（同じホームランでも構成がそれぞれ異なるので、HMMを用いる）。認識も音声認識と同様、インデキシングを行いたい動画から特徴ベクトルをつくり、それと学習でつくられたHMMを用いて認識を行う。

この手法を用いて、メジャーリーグの野球動画（ダイジェストデータ）で認識実験を行った結果、認識率は76.8％になったという。従来、このようなインデキシングは多大なコストをかけて、手動で行っていたことを考えれば、これは大きな進歩である。また、現在はシーンコンテキストの考慮やn-gramモデルなどを用いた研究を行っている。さらに、認識に音声情報を同時に用いるなどということも視野に入れて、認識率の向上を目指しているという。今後の更なる進展が期待できるところである。

さらに、こういった動画の認識において、人間の動作の推定・分析も研究の対象となっている。これはテレビ番組の検索に限らず、防犯カメラにおける不審者の察知や歩き方からの人物特定、お年寄りの安全確保など幅広く応用していくことができる。

そうだ、このDVDレコーダーには最新の動画認識・検索システムが搭載されているではないか。これであの選手のホームランシーンもすぐに見られるぞ。
ピンポーン、来客のお知らせです、カメラ認証中・・・、友人の○○さんです。

こんな便利で安全な未来のために、今日も研究が成されている。

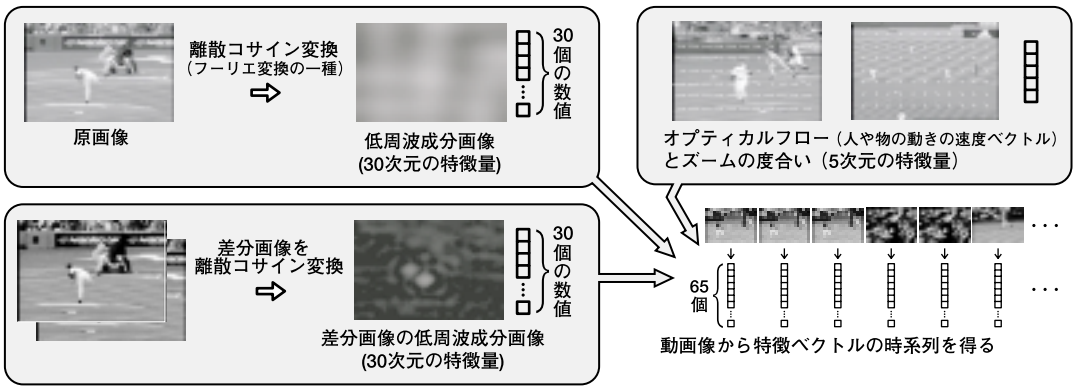


図６　動画の特徴ベクトル

今回は音声認識と動画像認識について紹介しましたが、古井・篠田研究室ではこの他にも対話システムや音声自動要約など様々な研究が成されており、誌面上の都合で全て紹介しきれなかったのが残念です。

また、取材では音声認識を利用した店舗情報検索システムのデモンストレーションを見せていただくなど、親切に対応していただきまして、ありがとうございました。最後になりましたが、大変お忙しい中、取材に応じてくださった先生方、研究室の皆様に深くお礼申し上げます。

（井上　中順）