論文の内容の要旨

論文題目　Japanese Prosody Analysis and its Application to Computer-Aided Language Learning Systems

和訳　（日本語の韻律解析と発音教育システムへの応用）

氏名　石井カルロス寿憲

韻律の発声を支援する発音教育システムでは母語話者の発声とマッチングしてその誤差を評価するという手法が多く、その言語の韻律的特徴を人間の知覚の観点から考慮したうえで評価する手法は少ない。本研究では、人間が行うのと同様に韻律的特徴を抽出することを目指して、母語話者がいかに韻律を生成して、いかにその生成されたものを知覚しているのか、というような観点から、日本語の特殊拍、モーラ・リズム、アクセントとイントネーションについて分析を行った。

特殊拍

日本語の特殊拍（長音、促音、撥音）の誤った発声はコミュニケーションにおいて誤解をまねく恐れがあるので、学習者に対しては発音教育の重要な課題となる。

特殊拍は主に持続時間によってその音韻性が判別されるが、話速が変わると当然音の持続時間は伸縮するし、その伸縮率は音によって異なるという問題点がある。過去の研究では話速を考慮しない手法が提案されているが、ここでは話速に依存しない特殊拍の判別を目的とする。

まず話速によっていかに各音の持続時間が変化しているのかを調べた。音声資料としは、メトロノーム・ビートによって話速を制御して発声したものを用いた。メトロノームのビートを利用して話速を定量化し、話速の逆数（時間単位）と各音の持続時間との関係を調べた結果、線形回帰直線で数式化可能と判断した。各音素でも前後音によって持続時間が変動するので、このような文脈を考慮したtriphoneを扱うこととした。しかし、全triphoneの数式を求めるためには非常に大量のデータ処理が必要となるので、先行音と後続音の影響を別々に調べ、これらの影響を線形的に扱うこととした。

これらの数式を利用して、発声内容を既知とした場合、その発話区分の話速を推定する手法を提案

した。観測した発話区分の持続時間をその発話区分を構成する各 triphone の持続時間数式の和と定義し、この等式を解くことで話速を求めるという手法である。推定した話速の値から逆に各 triphone の持続時間の予測値を推定することも出来る。

このような数式関係を利用して、話速を考慮した特殊拍判別手法を提案した。この手法は、目標となっている音節が特殊拍を含むか否かの２つの仮定に基づく。各仮定において、話速を推定し、その話速値から目標音の持続時間を予測し、予測値と観測値との距離を求める。距離の小さい方を選択するという手法である。

この手法によって、よい判別率で話速無依存の特殊拍判別が得られた。比較対象として、発話区分の持続時間をモーラ数で割るという話速推定を用いた結果、提案した手法の優れた性能が示された。

モーラ・リズム

前課題の特殊拍判別とも関連があるが、各音素の長さを制御することでリズムが成立され、リズムが崩れた発声は不自然となるので、リズムの発音教育も重要である。ここでは音声信号からリズム・パターンを抽出することを目的とする。

日本語はモーラリズム言語と言われ、モーラ単位で規則的に構成されて発声されたものが、聞き手にモーラ等時性の感覚を与えることを示す。このような観点から音の持続時間に関して多くの研究がされて来たが、モーラ等時性の感覚を与えるにも関わらず、音響的に測定されるモーラ長には明確な等時性が現れないことが示されている。本研究では、モーラ等時性が音響的な面からいかに観測可能かを調べた。

実験では等時性の基本となるメトロノーム・ビートに合わせて発声させたものを分析した。音響的に観測可能な３つの点（母音開始、子音開始、破裂開始）とメトロノーム・ビートとの距離を測った。その結果、ビートの位置が子音開始よりも、破裂・母音開始に近かったことが観測された。つまり、破裂・母音開始のパワーの立ち上がり部分がリズム感覚に重要と考えられる。なお、等時性はモーラの基本構造である CV よりも、VC 単位で実現されていることが言える。ただし、摩擦音の場合は、ビート位置が母音開始には近いが、多少の距離が観測される場合があった。この場合は、他の動的な音響的特徴も考慮すべきである。

ピッチ・アクセント

日本語には個々の単語は固有のアクセント・パターンを有し、同じ音素系列の単語であってもアクセントにより異なった意味を持つので、これらの発音教育は重要である。過去の研究では単語のアクセント学習システムが開発されているが、本研究では文レベルの発音教育を目指し、アクセント句のアクセント型判別を目的とする。

本手法の特徴は、F0 パターンをそのまま扱うのではなく、モーラ単位ごとに代表的な F0 値（以降、F0mora）を抽出することである。単語のアクセント型モデルでは F0 の平均値（avg）が F0mora として扱われたが、特に１型の場合、第２モーラの F0mora が、第１モーラのよりも高く抽出されていた。そこで、人間が知覚するピッチはモーラの終端側のターゲット値ではないかという仮説をたて、F0mora の候補として、１次回帰分析によって求められる F0 のターゲット値（tgt）も考慮した。ま

た、リズムもアクセントの知覚に影響する可能性も考慮して、CV単位とVC単位についても検討した。なお、モーラごとのピッチの動きをモデル化するため、隣接モーラの対数FOmoraの差として、FOratioという変数を定義した。

分析用のデータベースとして、人間によって各アクセント句のアクセント型がラベル付けされたものを用いた。各FOmora候補において、アクセント型別にFOratioの系列の分布を求めた。これらの分布の視察により、全体的にはFOmora(tgt;VC)が1型と2型の判別において、最も人間の知覚に近い結果を示した。

求められた分布を利用して、アクセント型とアクセント句長別にGaussian Mixture Model (GMM)を構築した。モデル学習としてデータベースの1部を使用し、残りは評価用に使用した。人間によってラベル化されたアクセント型を正解としているので、システムによる認識率は人間の知覚との一致率に対応する。結果としては、FOmora(tgt;CV)とFOmora(avg;VC)が78%程度で最もよい認識率を示している。期待されていたFOmora(tgt;VC)は67%の認識率となったが、これはターゲット値の求め方に問題があると考えられ、1次回帰以外の方法も検討すべきだと考えられる。また、アクセント句の直前のF0値も適切にモデルに追加すれば、認識率は向上すると考えられる。これらは今後の課題として残される。

イントネーション

イントネーションは統語や談話といった情報の伝達のみならず、発話の様式によって変化し、意図の伝達にも重要である。日本語では特に文末・句末でのイントネーションによって、文の意味・役割が変化し、礼を失した表現になる場合もあるので正しいイントネーションの発音教育も意義ある課題である。

本研究では6種類のイントネーションについて、文末の音響的特徴と知覚された印象との関連を調べた。音響的特徴としては、文末の持続時間(dur)、平均モーラ長(mora_dur)、先行フレーズにおけるモーラ数(rel_dur)、F0の傾き(F0_s)、パワーの傾き(pow_s)、と文末内F0ターゲットの差分(dF0_t)を用いた。

音声資料としては、外国人のための日本語教材に付属した音声を使用して、まず、母語話者が6種類のイントネーションを分類出来るかを調べ、83%の識別率が得られた。これは、異なった種類が類似した特徴で実現される可能性を示している。

被験者同士の結果が一致したものを分析に用いた。各音響的特徴において、イントネーションの種類別に分布を求め、これらを利用したGMMを構築した。同じデータベースで認識を行った結果、80%の認識率が得られた。

次に、これらの音響的特徴と、人間が知覚可能な特徴との対応付けを行った。ここでは、文脈の影響を防ぐため、非母語話者によって実験を行った。文末のトーンと、先行フレーズに対する強さ、長さ感覚をいくつかの段階のいずれに聞こえるかを判断させた。結果としは、文末トーンはdF0_t変数と、そして長さ感覚はrel_dur変数とよい関連性を示した。しかし、強さ感覚においては、より適切な相対的パワーの表現が必要である。なお、このような知覚的特徴との対応付けは、発音教育システムにおいて、学習者が知覚的に理解出来るフィードバックとして重要だと考えられる。

ピッチ知覚

前章では、アクセントやイントネーションの知覚において、いくつかの仮説のもとで適当な音響的特徴を提案した。本研究では、アクセント・イントネーションのパターンをより適切に表現するため、ピッチ知覚における音響的特徴との対応を詳細に調べた。

ここでは、FOmora の候補として、(avg/tgt;VC/CV) の他にもパワーによった重み付け (weighted/non-weighted) も考慮した。これは、音声のより強い部分がピッチ知覚に重要だという仮定に基づいている。

ピッチ知覚を調べる実験として、楽器音 (MIDI) のピッチが半音より細かい段階で調整出来るツールを作成し、自然音声から切り出した音節区間を被験者に聞かせて、知覚されたピッチに楽器音のピッチを合わせるよう指示した。被験者同士の知覚したピッチの平均値を人間が知覚した正解ピッチ (FOhuman) として扱った。そして、FOhuman と音響的に求められた FOmora の候補 (avg/tgt;VC/CV;w/nw) との偏差を調べた。その結果、FOmora(X,X,nw) に対し、FOmora(X,X,w) の方が小さい偏差を示した。全体的には FOmora(avg;VC;w) と FOmora(tgt;CV;w) が最もよい結果を示した。この結果はアクセント型判別タスクで得られた結果と一致し、人間の知覚により近いパラメータを用いてモデル化すれば、認識率も向上すると言える。

個々の音節についても詳細な分析を行った結果、ピッチの変動によって異なったパラメータがより FOhuman に近い傾向が見られた。従って、パラメータの組み合わせにより、人間が知覚するピッチにより近い表現が求められると考えられる。

結論

本研究は日本語の韻律の発音教育システムを構築するため、音声生成や音声知覚の面からさまざまな音響的特徴を調べ、言語的情報との対応をモデル化する手法を提案した。なお、この論文で得られた結果は何らかの加工により、発音教育システムのみならず、音声認識を利用した多くの音声理解や音声対話システムに生かせると考えられる。