論文の内容の要旨

論文題目
データベースに基づくテキストからの基本周波数パターンの生成
(Data-Driven Generation of Fundamental Frequency Contours from Text)

氏名　　桜井 淳宏

概要

テキスト音声変換（TTS）における難題の一つは韻律的特徴、特に基本周波数（F_0）パターンの生成である。従来の TTS システムはルールを用いて F_0 パターンの生成を行なってきたが、韻律的特徴の生成ルールにはヒューリスティックスに依存する部分が大きく、最近はそれに対してデータベースの学習に基づく手法（いわゆるデータドリブン手法）を導入する傾向が見られる。データドリブン手法とは、F_0 パターンの生成に必要な言語情報とそれに対応する F_0 パターンの例をデータベース（以下韻律データベース）として収集し、両者の量的な関係を自動的に学習しようとする手法のことである。しかしながら、韻律的特徴には本質的に様々な要因による変動があり、テキストから抽出可能な言語情報と韻律的特徴との関係を導出することは難しい。よって、提案されているデータドリブン手法のほとんどはその関係をうまくとらえることができない。

このような観点から、データドリブン手法を実現するためには、何らかの制限を導入することによって学習効率を上げる他、大量のデータベースの作成を可能にする手段や学習方式を開発する必要がある。

そこで、本研究では、データベースに基づく F_0 パターンの生成を実現すべく、いくつかの手法やアルゴリズムを提案する。まず、前半には韻律データベースの様式やその作成に役立つ一連のアルゴリズムを提案する。提案している韻律データベースの大きな特徴は F_0 パターンの生成過程モデル（以下 F_0 モデル）のパラメータを利用することである。F_0 モデルは少ないパラメータで基本周波数パターンを表現できるため、より効率よい学習が可能になる。一方、韻律データベースを作成する際、問題となるのは大量のデータに対応する F_0 モデルのパラメータ値を推定するために必要な労力を減らすことであるが、F_0 モデルのパラメータ値を自動的に（若しくは半自動的に）推定できることが望ましい。そこで、この問題を解決するためのアルゴリズムを提案する。

後半には学習・生成問題に焦点をあて、上記の韻律データベースを利用して F_0 パターンの

生成を実現するための手法を提案する。その一つはニューラルネットワーク或いは２分木を用いた推定モジュールにより、F_0モデルのパラメータを推定し、推定したパラメータをもとに F_0パターンを生成する手法である。もう一つは、モーラ単位の F_0パターンの形状と平均値をコードブック化し、それらのコードを出力する離散HMMを用いて F_0パターンをモデル化・生成する手法である。以下、データベース作成の段階から研究の内容をより詳しく説明する。

韻律データベースの作成に関する研究

データドリブン方式の学習効率をあげる方法として、F_0パターンを直接表現する代りに F_0モデルのパラメータを用いることが考えられる。そこで、F_0モデルのパラメータを含む韻律データベースの仕様や作成について検討し、F_0モデルのパラメータを用いて半自動的にラベリングするための手法を提案する。この手法は、F_0パターンの形状から韻律境界（フレーズ境界およびアクセント境界）を別々に抽出することに基づく。フレーズ境界とアクセント境界はそれぞれ F_0モデルのフレーズ指令とアクセント指令に対応する境界である。次に、フレーズ・アクセント境界やその他種々の言語情報（主に文節境界等）を用いて F_0モデルの指令の発生時刻の初期値を決める。最後に、実測の基本周波数パターンを基準とした AbS（合成による分析）処理を行ない、すべてのパラメータ（指令の大きさと発生時刻）の微調整をする。

以上のラベリング手法を評価した結果、人間による修正作業と組み合わせれば効率良くデータベースのラベリングを行なうことができるほか、言語情報（品詞、アクセント型等）を更に導入することによって自動推定の精度が向上することがわかった。そこで、ラベリング過程において言語情報を積極的に利用する試みとして、複合名詞のみの F_0モデルパラメータを自動的に求める手法を提案する。

ニューラルネットワーク及び２分木による F_0モデルパラメータの推定

学習用の韻律データベースを作成した後、次はデータベースからの学習に基づく F_0パターンの生成手法について検討する。ここでは、ニューラルネットワーク或いは２分木に基づく推定アルゴリズムを用いて、テキストから抽出される言語情報をもとに F_0モデルのパラメータを推定する手法を提案する。推定の基本単位は韻律語（一つのアクセント指令に相当する文のかたまり）とし、その言語的な属性と F_0モデルパラメータとの量的な関係をデータベースから学習する。学習後は、テキストから抽出した韻律語及びその言語的属性のみを入力として F_0モデルパラメータを推定し、そこから F_0パターンを生成する。ただし、テキストの形態素解析処理は既に行なわれているものとする他、音素の持続時間も既知であると仮定する。入力に用いる韻律語の属性としてはその位置、アクセント型および構成単語数、更に単語の品詞や活用に関する情報を利用する。推定する特徴量は F_0モデルパラメータのタイミングや大きさである。

最初に提案する学習手法はニューラルネットワークに基づくものである。ニューラルネットワークは非線形な問題に向いているため、言語情報から F_0 モデルパラメータへのマッピングに適切であると思われる。ここでは３種類のニューラルネットワーク構造（エルマン型、ジョルダン型、多層パーセプトロン型）を利用し、それぞれによる結果を比較する。多層パーセプトロンは最も一般的に応用される構造である。エルマン型とジョルダン型はいずれも再帰構造であるが、エルマン型は隠れ層から、ジョルダン型は出力層からのフィードバックを有する。中国語やドイツ語において、フィードバックを有するニューラルネットワークを用いて F_0 パターンの生成を行なおうとする研究例はあるが、そこでは音節単位で F_0 パターンを処理しているため、フィードバックはパターンの連続性を維持する役割を果たしている。一方、本手法では連続性は F_0 モデルによって自動的に確保されているため、フィードバックは単に前後の韻律語の影響を表そうとしている。

ニューラルネットワークによる手法を評価するため、自然音声から抽出した F_0 パターンとの平均自乗誤差（MSE 誤差）を求める。実験では構造や隠れ層の要素の数による変動がそれほど見られなかったが、１０要素の隠れ層をもつエルマン型ネットワークによるＭＳＥ誤差が最小となった。

一方、もう一つの学習方式として２分木を用いた手法がある。ここでは、ニューラルネットワークによる手法と比較するために２分木に基づく推定モジュールを構築、評価する。２分木によってモデル化できる問題の種類は限られているが、学習の結果として得られる知識の可視性に関してはニューラルネットワークより優れている。推定した F_0 モデルパラメータを用いて F_0 パターンを生成し、自然音声から抽出したパターンとのＭＳＥ誤差を求めた結果、ニューラルネットワークによる結果とほぼ同程度であることがわかった。更に、ニューラルネットワーク方式で得られた代表的な F_0 パターンと比較するために簡単な聴取実験を行ない、ニューラルネットワークの方が多少高い評価を得た。最後に、従来のルールに基づく手法と比べても、両手法とも比較的良い品質の F_0 パターンを生成できることがわかった。

モーラ遷移に基づく離散HMMによる F_0 パターンのモデル化と生成

以上述べた手法はいずれも F_0 モデルパラメータに基づくものであるが、F_0 モデルパラメータが付与されたデータベースの存在が必須となる。一方、F_0 モデルパラメータのラベリングを必要としない方法として、モーラ F_0 パターンのクラスタリングと離散隠れマルコフモデル（HMM）に基づく F_0 パターンのモデル化・生成手法を提案し、その有効性について検討する。

この手法は、韻律語を一つの HMM でモデル化し、その状態遷移をモーラ遷移と対応付ける。HMM の出力は２次元ベクトルであり、一つのコードはモーラ単位の F_0 パターンを近似的に表し、もう一つはモーラ単位の平均 F_0 の差分を離散化したものである。HMM の学習は通常の音声認識アルゴリズムと同様の方法で行なう。

学習を行なった後、HMM から F_0 パターンを生成するが、Viterbi アルゴリズムに基づく方法を用いる。Viterbi アルゴリズムは本来、音声認識において、任意な出力符号系列に対してその尤度及び生成過程に対応する最適パスを推定するために用いられる。一方、HMM から出力系列を生成したい場合、出力符号系列の長さは既知であるものの、出力符号系列そのものが与えられていないため、Viterbi アルゴリズムをそのまま利用できない。ここでは Viterbi アルゴリズムで用いられる距離関数を変更することによって最適パス及び最適出力ベクトル系列を求め、それを用いて F_0 パターンを生成する。

評価実験によると、アクセント核の位置等、韻律的特徴をモデル化することができるが、F_0 モデルを用いた手法と比較すると、必要とする学習データが大きいことがわかった。一方、F_0 モデルパラメータを含むデータベースを必要としないため、データベース作成の自動化をはかれる。