論文の内容の要旨

論文題目　　音声の基本周波数パターン生成過程モデルの特徴パラメータ自動抽出手法とそれを用いたコーパスベース韻律生成

氏　名　　成澤　修一

近年，計算機技術の急速な発展と情報ネットワークの広汎な拡大・普及に伴い，機械により処理・蓄積された膨大な量の情報を常時・至る所で利用することが可能になりつつある．それに伴って，これらの情報を利用する人間との間に情報の迅速・円滑な授受がますます必要とされるようになった．特に音声言語は，人間同士の情報の授受において最も容易で迅速な媒体であるため，これを人間と機械の間の情報交換手段として活用するために，現在，機械からの音声言語の出力(音声合成)や機械への音声言語による入力(音声認識)の技術が鋭意研究され，徐々に実用化に向かっている．しかしながら従来の音声言語入出力技術は，韻律的特徴をほとんど利用していなかったため，十分に高度な性能を達成し得なかった．

音声の韻律的特徴は，本来文字言語にも含まれる語義・統語・意味・談話などの情報(言語情報)のみならず，文字言語には陽に含まれない話者の意図や態度に関する情報(パラ言語情報)や，話者の個体差，性別，年齢，感情などに関する情報(非言語情報)をも含んでいる．したがって，高度な音声入出力技術の実現には，まず，これらの情報と韻律的特徴との関係についての知識を獲得することが不可欠であり，そのためには，大量の音声データの統計的な性質を利用する分析手法が有力である．しかしながら，音声の韻律的特徴は，例えば基本周波数の時間的変化のパターン(F_0 パターン)のように，大量のコーパスを音声波形やスペクトルといった次元でそのまま用いるのでは正確には表現できないものである．したがって，それらを生成する段階，つまり，対象とする現象をモデル化し，そのモデルパラメータの次元で表現することが必要となる．本論文では，このような見地から，F_0 パターンを生成するモデルとそのパラメータに着目し，音声信号からそのパラメータを自動的に抽出する手法と，それを音声言語情報処理に利用する手法を検討した．

音声の韻律的特徴を表現する主要な物理量としては，声帯振動の基本周波数(F_0)，単音の持続時間長，音源の強度が挙げられる．日本語をはじめ多くの言語では，特に F_0 パターンが構文や意味の伝達に重要な役割を果たし，藤崎らによる生成過程モデルを用いれば，少数のパラメータ(フレーズ指令およびアクセント指令の生起時点とそれらの大きさ)によってその特徴を正確・定量的に表現しうることが広く知られている．しかしながら，観測された F_0 パターンからモデルのパラメータを抽出することは，いわゆる逆問題であって，この場合には解析的に解くことはできず，モデルのパラメータの初期値を出発点とした逐次近似を必要とする．この場合，高精度のパラメータを迅速に抽出するには適切な初期値の設定が不可欠であるが，従来はこれを人手によって行っていたため，大量の音声資料の自動的処理は困難であった．もし，適切な初期値を自動的に求めることが可能になれば，パラメータ抽出にかかる労力や時間が軽減されるだけでなく，大量の韻律パラメータが得られるため，生成過程モデルに基づく大規模な韻律コーパスの構築も可能になる．したがって，実測の F_0 パターンからパラメータの初期値を自動的に決定し，さらにそれに基づいて高精度のパラメータ抽出を自動的に行う処理が必要である．

実測の F_0 パターンには，F_0 の抽出誤りや子音調音による F_0 の乱れなど，生成過程モデルで考慮していない現象や，F_0 の存在しない無声・無音区間が含まれているため，パラメータ抽出に先立ち，これらを修正・除去・補間することが重要である．一方，フレーズ成分の時間的変化はアクセント成分よりもはるかに緩やかであるため，F_0 パターンの変曲点の位置はアクセント成分のそれとほぼ一致する．したがって，実測の F_0 パターンから上述の変動要因の影響を除去する前処理を行ったうえで，それを至るところ連続かつ微分可能な 3 次曲線で区分的に近似すれば，その変曲点の位置は，その曲線の 1 次導関数の極値の位置として，容易に求めることができる．このような近似を行うことにより，解析的には解けなかった逆問題が 1 次方程式の解を求める問題に帰着される．

以上の観点から，本論文では，実測された F_0 パターンを至るところで連続かつ微分可能な曲線によって近似するための処理(前処理)，得られた曲線からアクセント指令とフレーズ指令のパラメータの初期値を決定するための処理(初期値抽出処理)，さらに，それらの初期値をもとに逐次近似によりパラメータの最適値を求めるための処理(最適化処理)，の 3 段階の処理からなる手法を提案した．

東京方言話者の日本語朗読音声として，NHK85 文(書物の 1 章の読み上げ．男声話者 1 名: 資料 A-1，女性話者 1 名: 資料 A-2)と ATR503 文(個々には無関係な文．男声話者 1 名: 資料 B)を対象として分析を行った．韻律研究に深く携わっているエキスパートが抽出した指令を正解として欠落誤り率・挿入誤り率を算出した結果，フレーズ指令とアクセント指令について，資料 A-1 では，11.3%と 7.6%，19.7%と 13.3%，資料 A-2 では，14.0%と 9.3%，32.5%と 30.0%，資料 B では，6.1%と 16.5%，16.2%と 7.8%，であった．また，この手法の適用性を日本語以外の言語の F_0 パターンについても検証するため，英語 50 文(男性 2 名，女性 2 名)とポルトガル語 3 文（ただし各 5 回の発話．男性 3 名，女性 2 名)の音声資料を対象とした分析を行った結果，指令抽出に関する欠落誤り率・挿入誤り率は，英語音声資料では， フレーズ指令で 35.7%と 15.5%，アクセント指令で 14.5%と 17.5%，であった．

なお，上記の方法を用いれば，生成過程モデルのパラメータを自動的に抽出する事が可能であるが，音声合成・認識への応用を考えた場合，テキストから得られる言語情報と対応の取れた指令をいかにして推定するかが重要である．したがってさらに，日本語を対象として，テキストから得られる文節の係り受け情報と語のアクセント型に関する情報を，上述の手法に組みこむ方法を提案した．前述の日本語音声資料 A-1 を対象として実験を行った結果，言語情報を利用することにより，欠落誤り率・挿入誤り率が，フレーズ指令に関してはそれぞれ 1.1%と 0.7%，アクセント指令に関してはそれぞれ 7.7%と 2.6%減少することを確認した．

本研究で提案した生成過程モデルパラメータ自動抽出手法は，韻律生成をはじめとして，韻律構造の推定，感情の推定など，音声言語処理における多くの課題に適用できるが，本論文では，これを用いて韻律情報を付与した音声コーパスを作成し, さらにそのコーパスを用いて F_0 パターンの合成システムを構築した．

現在，全世界で多数の大規模な音声コーパスが構築されているが，それらのコーパスのほとんどは分節的特徴の利用を目的としたものであり，韻律的特徴に関する記述のなされているものは比較的少数である．また，韻律的特徴に関する記述のなされた音声コーパスの多くは，英語の韻律を記述する ToBI や，そこから日本語用に拡張された J-ToBI や X-JToBI などの枠組みの上で構築されている．しかし，これらの方法による韻律の記述は，韻律的特徴の物理的な性質に基づく客観的・定量的なものではなく，ラベラの主観に基づく定性的なものである．また，人手によるラベリングは非常に時間がかかる作業でもあるため，自動ラベリングを行う試みもなされてきたが，必ずしもその結果は満足の得られるものとはなっていない．これに対して，本論文で提案した生成過程モデルパラメータ自動抽出手法を用いれば，韻律的特徴に関する定量的な記述のなされた音声コーパスの構築が可能であり，しかも生成過程モデルのパラメータを利用することにより，限られた量の音声コーパスからでも破綻の少ない韻律生成の可能な統計モデルを構築しうるものと考えられる．

この考えに基づいて，まず，既存の音声コーパスに対して，各発話ごとに音声信号から上記の手法により F_0 パターン生成過程モデルのパラメータを抽出すると同時に, 既存の音声認識ツールを用いて単音の認識とその境界の決定を行う．一方，その発話と対応する漢字仮名混じり文を既存の統語解析ツールを用いて解析し，形態素・統語構造・係り受けの決定を行うとともに，音声言語情報と文字言語情報との相対時間関係を決定し，韻律コーパスを構築した．また，このコーパスを韻律の規則合成に利用するため, 生成過程モデルの指令の生起時点・大きさを指定する決定木を, 既存のツールを用いて作成した．次に，この韻律コーパスを用いたテキストからの音声合成システムを構築した．このシステムでは，与えられた漢字仮名混じり文を前記の統語解析ツールを用いて解析し，形態素・統語構造・係り受けの決定を行うとともに, 分節的特徴に関しては隠れマルコフモデル(HMM)を利用してメルケプストラムの時系列を生成し，韻律的特徴に関しては上記の決定木を用いて生成過程モデルの指令の生起時点・大きさを指定し，音声合成を行う．実測の F_0 パターンを構築した韻律コーパスから得られる F_0 パターンにより置き換えた分析再合成音と，上記のシステムによる合成音声とに対して，それぞれ韻律に着目した 5 段階の主観評価実験を行った結果，本手法により作成した韻律コーパスの有効性を確認し得た．