重畳モデルと声調核モデルに基づく柔軟な韻律制御手法と

それによる標準中国語音声合成システム

孫　慶華

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り、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)などの音声情報処理技術が鋭意研究され、徐々に実用化に向かっている。

そして、音声合成に関しては、波形選択合成方式や HMM 音声合成方式の導入等により、合成音声の品質及び柔軟性は格段に向上したが、イントネーションやリズム等の韻律的特徴の制御には、なお問題が多い。特に、典型的な声調言語として知られている標準中国語（以下、単に中国語と記述）では、声調が意味の区別等、コミュニケーションの実現に大変重要な役割を果たしており、その適切な制御が重要である。しかし、一般的に声調言語では非声調言語と比べ、基本周波数（F0）の変化が激しく、その制御はより困難な課題となる。

最近、大量の音声データベースの整備と、計算機によるデータ処理能力の向上を背景に、コーパスベース手法における F0 制御に関する研究は盛んになってきた。従来の規則に基づいた F0 パターン生成方式（ルールベース手法）の多くが発見的な手法に基づいているのに対して、コーパスベース手法は大量のデータを用いた自動学習や単位パターン選択に基づいているため、高品質で自然性の高い F0 を合成しやすい、というだけでなく、システムの自動学習が可能、音声データ提供話者の個人性、更には発話様式が合成した F0 パターンによく反映される、などの特徴をもつ。そして、liner regression method, hidden Markov models (HMM), tree-based approach, neural network, vector quantization などの統計手法を用いる様々なコーパスベース手法が提案された。しかしながら、こういった手法はパラメータと言語情報の対応性が薄いため、大量な学習コーパスが必要となる。且つ、人間では学習したルール(特徴パラメータと言語情報との関連関係)を理解しがたいことが多く、一旦機械学習で得られるモデルを修正することが難しいため、コーパスに入っていない発話スタイルを再現する時、あらかじめ、そのようなコーパスを用意することが必要となる。そのため、コーパスベース手法では、柔軟な制御が困難である。その代表的な統計的手法として、隠れマルコフモデル(HMM)は、すでに音声認識及び音声合成において、一般的なアプローチとなってきた。他のコーパス手法と比べ、話者適用技術の開発により、少量データで目標話者の音声を合成することが実現できるが、韻律の制御においては、まだ色んな問題が残っている。前後の音素や当該音節が文中に占める位置などのいろんなコンテキスト情報を入力として学習に与えているが、単語やフレーズなどの長い単位での F0 の動き

を直接に考慮していないという問題がある。さらに、このような統計的手法では、過剰な平均化及び平滑化などの原因で、生成された F0 パターンが平ら（メリハリのない）になっていることが多い。

F0 パターンを表現するモデルの中に、長い韻律単位を考慮したモデルとして、F0 パターン生成過程モデル(以後 F0 モデル)が最も高く評価されている。このモデルは、人間の声帯の振動機構を定量化したもので、句頭から句末にかけて緩やかに降下するフレーズ成分と、局所的起伏するアクセント成分（中国語では声調成分）との和によって対数 F0 パターンを表すものである。日本語 F0 パターンの生成過程を模擬するものとして提案されたが、負のアクセント（声調）指令を導入し、正負のアクセント（声調）指令の対を声調指令とすることで、中国語にも対応するように拡張された。このモデルによって、中国語の F0 パターンを精度良く生成できることはすでに示されている。日本語においては、F0 モデルのパラメータであるフレーズ指令及びアクセント指令（中国語の場合は声調指令とも呼ばれる）は言語情報との関連性が大きいため、ルールベース手法での制御はすでに成功している。そして、パラメータの自動抽出手法も提案され、大規模データベースの構築を可能にすることで、コーパスベース手法の制御も成功している。しかし、中国語の声調成分は非常に複雑な特徴を有し、声調指令として、正のステップ関数、負のステップ関数、あるいはその組み合わせを考える必要があるため、適切な規則を発見するのは容易ではない。また、高精度なパラメータ自動抽出手法が樹立されていないので、声調指令情報を有する十分な量の韻律コーパスを用意するのも困難といわざるを得ない。従って、F0 モデルの枠組でのルールベース手法及びコーパスベース手法での中国語 F0 制御法に関しては、どちらも樹立されていないのが現状である。

以上のような観点から、我々はすでにルールベース手法とコーパスベース手法を融合した 2 段階 F0 パターン合成法を提案した。この方法では、F0 パターンをフレーズ成分と声調成分とに分けた上で（以下、重畳法)、規則でフレーズ成分を生成したあと、生成したフレーズ成分の情報をコーパスベース声調成分生成手法に反映させるものである。声調成分の生成は、F0 モデルの枠組みではなく、各音節の声調核部分の蓄積した F0 の声調成分パターンをコーパスから選択し、接続することによって行う。ここで“声調核”とは、中国語音節の F0 パターンの中で、相対的に前後のコンテキストに影響されにくく、声調本来の韻律的特徴を保持している部分を指している。 更に、中国語では日本語と違って、フレーズ成分と声調成分とかなり高い相関を持っていることから、フレーズ成分を生成した後、その情報を声調成分の推定に用いる 2 段階合成法も提案した。そして、TD-PSOLA を用いて、録音した音声の F0 パターンと合成したものと入れ替え、いくつかの音声再合成実験を行った結果、朗読音声について、提案し手法は十分に自然性の高い F0 パターンが生成できることが証明された。そして、HMM で生成した F0 パターンと比較し、提案手法の方がより自然性の高い F0 パターンが生成できることを示した。しかし、さらなる比較実験で、声調核モデルを用いれば、フレーズ成分及び声調成分に分割せず（以下、一括法）に、直接に F0 パターンを合成しても、同じく自然性の高い F0 パターンが生成できることが分かった。重畳法の優位性は、柔軟な韻律制御にあることを実証するために、強調したい単語の前に大きめのフレーズ指令を手作業で設定し、重畳法で F0 パターンを生成した。意図した

単語が正しく強調されたと知覚されるかを聴取実験によって、強調音声をはっきりラベリングしたデータベースを用いなくても、ある程度単語の強調が実現できた。このような方法で完全な強調音声の制御ができると言い難いが、提案手法は十分な高品質を保ちながら、韻律制御の自由度（柔軟性）の高い手法であることが証明できた。

応用として、本研究で提案した F0 パターン生成法を用いて F0 パターンを生成し、音声素片接続型合成システムに素片選択の目標値として利用したり、発話スタイルの修正に目標 F0 パターンとして用いたり、CALL システムに正解の教師データとして学習者に呈示したりするなどいろんな用途が考えられる。

今まで行った F0 パターン生成実験では、なるべく合成音声に F0 パターン以外の要素に影響しない様、録音した音声の F0 パターンのみを修正し、音声を再合成したが、実際に合成システムに適用するという観点から見た場合、提案の F0 パターンの有効性を検討するために、それを音声合成システムに組み込み、音声合成実験を行う必要がある。更に、継続長やパワーなど他の韻律特徴量の制御手法を検討するためのプラットフォームとしても、音声合成システムの構築が必要であると考えられている。従って本研究では、標準中国語音声合成システムを構築した。

本研究で構築した音声合成システムはテキスト処理部と音声合成部で構成されている。テキスト処理部では、テキストを入力として形態素解析、構文分析、韻律階層アライメントなどのプロセスが行われ、漢字の読み、声調、及び単語の品詞情報、各韻律階層の境界、ショートポーズの挿入位置などが決定される。そして、音声合成部には、継続長予測モジュール、F0 パターン生成モジュール、スペクトル生成モジュール及び音声波形合成モジュールなどが含まれている。継続長予測モジュールでは、音節を音素に分割し、各音素の継続長を 2 分岐決定木によって予測する。その入力としては、テキスト処理部で得られた分析結果を用いる。これらの継続長情報は F0 パターン生成モジュール及びスペクトル生成モジュールに渡され、後続のプロセスに用いられる。F0 パターン生成モジュールでは、本研究で提案した手法を用いて、ルールベース手法によって音節単位で生成した声調成分を、ルールベース手法で生成したフレーズ成分に乗せ、F0 パターンを出力する。そして、スペクトル生成モジュールは Single Gaussian output distributions を持つ音素単位の left-to-right 型の 5 状態 HMM を用いて構築される。スペクトル及びピッチの静的特徴量（24 次元の mel-cepstrum 係数及び F0 値）とそれぞれの動的特徴量（デルタ及びデルタデルタ値）の結合したものを特徴ベクトルとし、mel-cepstrum 係数は連続分布、F0 値は多空間上の確率分布(MSD)で HMM を定義した。合成するとき、先行プロセスで予測した音素継続長を与えて、音素境界を固定し、フレーズごとに mel-cepstrum 係数及び F0 値を出力する。そのため、一般に使われている連続学習ではなく、音素境界を固定した学習手法を用いた。ここで、出力した F0 値は単なる有声／無声の判定に用いることにした。最後に、F0 パターン生成モジュールで生成した F0 パターンとスペクトル生成モジュールで出力した mel-cepstrum 係数の時系列を MLSA フィルターに入力し、音声波形を生成する。構築した合成システムを稼動し、音質の高い音声合成が実現できることを証明した。今後、より柔軟性の高い合成システムの構築を目指して、F0 パターン、継続長及びパワーの制御の一貫した手法を開発する予定である。