高品質音声合成のための韻律制御[#]

Prosody Control for High-Quality Speech Synthesis

東京大学大学院新領域創成科学研究科

School of Frontier Sciences, University of Tokyo

広瀬　啓吉

Keikichi Hirose

The aim the research group is to establish an advanced scheme of controlling prosodic features for speech synthesis, enabling to generate synthetic speech highly human-like in various utterance styles. Scope of the study is not only limited to realizing linguistic information, but also to including para- and non- linguistic information in synthetic speech. Realization of a spoken dialogue system, whose reply speech is highly acceptable for users, is also aimed at. Although both of heuristic and statistical frameworks are introduced for the research works, major focuses are place on two corpus-based methods for F_0 contour generation; HMM-based method and method based on the generation process model. Roles of non-lexical utterances in dialogues are investigated through analysis of their prosodic features. The research also covers the control of prosodic features in the automatic generation of presentation contents, which includes linguistic and non-linguistic modalities, and in multi-modal interface with character agents. Through four years of research work, a number of outstanding results have been accomplished; corpus based generation of F0 contours for emotional (including calm) speech synthesis using automatically generated prosodic corpus in the framework of the generation process model, role of prosodic features in conveying information in non-linguistic utterances, synchronized control of face and voice to emphasize important parts of automatically generated presentation contents, multi-modal agents possible to express their emotional state visually and acoustically, HMM speech synthesis with various voice qualities and utterance styles based on average voice models and speaker adaptation, unified HMM speech synthesizer and realization of emotional speech, and so on.

Keywords: Speech Synthesis, Prosody, Generation Process Model, HMM Synthesis, Corpus-based Synthesis, Non-linguistic Utterance, Multi-modal, Emotional Speech, Presentation Content Generation

1. 研究の目的

最近のテキストから音声を合成する技術の進展は著しく、特に、音声コーパスから統一的な評価基準の基に音声素片を切り出して接続するコーパスベース合成等によって、合成音声の品質が一段と向上した。しかしながら、これを人間と機械との間の円滑な情報授受を実現するための音声合成技術としてみた場合には、問題が多い。すなわち、現在の音声合成システムの多くは、単にテキストを（単語のアクセントをつけて）棒読みするもので、韻律の観点からみると、文の意味や内容を的確に表現するものとなっていないという問題がある。また、韻律は、言語情報のみならず、パラ言語情報（意図・態度）、非言語情報（感情）の伝達に重要な役割を果たしているが、これについては一切考慮されていない。この様な観点から、本研究では、意図・態度・感情等の伝達をも視野に入れた上で、朗読調のみならず、対話調等の種々の調子の音声を、従来になく人間らしい抑揚で合成する技術を確立することを目的とする。また、成果をもとにユーザフレンドリ

[#] 研究課題：高品質音声合成のための韻律制御，研究課題番号：12132202

な応答音声生成システムを構築する。これによって、人間にとって聞きやすくかつ分かりやすい音声の合成が可能となる。この様な目的を達成するために、対話音声を中心とした連続音声の韻律的特徴を定量的に分析し、談話情報との対応を明らかにするとともに、言語行動学的側面からの意味付けを試みる。また、意図・態度・感情等の言語情報以外の情報も対象とした韻律との対応づけを試みる。従来この様な対応関係の構築は人間が発見的に行うことが多かったが、本研究では統計的手法を駆使し、合成に有効な対応関係を構築する。本研究は、①代表的な韻律的特徴である基本周波数の時間変化パターン（基本周波数パターン, F0パターン）を、代表者らが提案した生成過程のモデルに基づいて分析し、特徴の明確な把握を可能とする、②対話音声を対象とし、従来のテキスト音声合成では取り扱われていない談話情報と韻律との関係を分析し、韻律合成規則に組み入れる、③相槌等の短文による談話情報等の有効な表現を可能とする、④生成過程のモデルの枠組みにおいて統計的手法を活用し、言語情報、パラ言語情報、非言語情報と韻律との関係を明らかにして韻律合成手法を確立する、⑤HMM（隠れマルコフモデル）に話者適応法を利用することにより合成音声の声質変換を実現する、⑥ヒューマン・マシン・システムの音声出力を想定し、文生成から音声合成への一貫した手法を構築する、といった特色・独創的な点を有する。

2. 研究の体制と成果

Table 1. Members and their research topics.

Name	Affiliation	Major research topic
Keikichi Hirose	University of Tokyo	Generation of prosody using generation process model
Toyoaki Nishida	University of Tokyo	Prosody control in automatically generated presentation contents
Nigel Ward (till 2003.3)	University of Tokyo	Prosodic features of grants in dialogue
Mitsuru Ishizuka (from 2003.4)	University of Tokyo	Prosodic control for multi-modal agent
Takao Kobayashi	Tokyo Institute of Technology	HMM-based speech synthesis with various styles of prosody
Keiichi Tokuda	Nagoya Institute of technology	HMM-based synthesis of emotional speech

Table 2. Group meetings.

Fiscal Year	Date	Place
2000	November 30	University of Tokyo
2001	July 3	
	January 11	
2002	November 28	
2003	July 4	
	December 17	

表1のように、本研究班は代表者広瀬と分担者西田、Ward（2003年3月まで）、石塚（2003年4月から）、小林、徳田で構成され、研究目的を達成するため、主として広瀬は生成過程のモデルに基づく韻律生成と応答文生成の観点から、西田は談話情報の抽出と韻律との対応の観点から、Wardは短文による談話情報、感性情報の表出の観点から、石塚はマルチモーダルエージェントによる感性情報の表出の観点から、小林と徳田は統計的モデル（HMM）による韻律合成の観点から、表2のように年2回程度のペースで班会合を行った他、電子メールを活用して、それぞれ緊密に連絡を取りながら研究を進めた。以下に、各成果の概略を述べる。

2.1 生成過程のモデルと統計的手法に基づくF0パターンの合成（広瀬）[1]

F0パターン生成過程のモデルの制約下で、F0パターンを合成するcorpus-based手法を開発した。これは、統計的手法によりF0を直接推定する代わりに、モデルのパラメータを推定するもので、モデルの制約のために、小さなコーパスを用いた場合でも聴覚上破綻の少ない韻律生成が可能となる。種々の発話スタイルを実現するに際して、各発話スタイルのモデルパラメータを付与した音声コーパスを用意する必要があるが、これを自動的に作成する手法も開発した[2]。

当初は、入力テキストをアクセント句単位に区分し、句単位ごとにモデルのパラメータを統計的手法（回帰木）により推定していたが、自動作成によって得た学習コーパスの不適切なモデルパラメータに起因する推定誤りが問題となった。学習コーパスのモデルパラメータの精度を向上するために、文節境界以外にはフレーズ指令は生起しない等の言語情報を用いた制約を加えたモデルパラメータ抽出手法を開発した。この手法に則した合成手法とするために、文節境界のフレーズ指令推定、韻律語（1つのアクセント成分からなる発話単位）境界推定（各形態素境界韻律語境界であるか否かを推定）、各韻律語のアクセント核決定、各韻律語の

アクセント指令推定の4つのプロセスからなる手法を開発した。1, 2, 4番目のプロセスは回帰木を用いた推定であるが、3番目のプロセスは構成単語のアクセント型とアクセント属性から規則によって行う。規則は研究室で別に作成したものを用いた。推定のための回帰木への入力は、該当及び先行単位の文法的情報、両単位間の構文的な境界の深さ、単位の長さや文内位置の情報等である。推定した指令の値（大きさと時点）を用いて生成した F0パターンと、自然音声の F0から抽出した指令値を用いて生成した F0パターンの誤差を評価したところ、新しい手法により大幅に改善することが示された。

同様な統計的手法により、音素長の推定を行い、得られた F0パターンと音素持続時間を用い、HMM 音声合成により感情音声（平静音声を含む）の合成実験を行った。合成音声の聴取実験の結果、怒りについては提案手法により良好に表現されることが示されたが、喜びと悲しみについてはさらに研究が必要なことが分かった。

今後は、平静音声と感情音声の韻律の違いを定量的に表現し、それによって、コーパスによらずに新しい話者の感情音声を平静音声から生成することを目指す。

2.2 言語・非言語モダリティを統合した会話エージェントのプレゼンテーション行動自動生成（西田）[3]

言語・非言語モダリティを統合した会話エージェントのプレゼンテーション行動自動生成に関わる一連の研究を行い、次のような成果を得た。

第一に、実データを分析して、対話テキストの内容を合成音声によって的確に表出する能力を持つ音声対話生成方式の基本設計を行った。読み上げ音声を聞くだけでは理解しにくい箇所を抽出して、言語処理・音声処理によって聞きやすく変換するアルゴリズムを実現することを目標に、話題や焦点などの対話内容の理解の手がかりとなる言語表現と韻律の関係の分析を行った。

第二に、自然言語テキストの談話構造解析に基づく韻律制御システムのアーキテクチャの設計を行った。この方式では、与えられた自然言語テキストを解析して、エージェントのジェスチャーや顔表情などの非言語的なモダリティを決定し、汎用的なエージェントアニメーション自動生成システム CAST (The Conversational Agent System for neTwork applications) における会話エージェントのプレゼンテーション行動を生成する。

第三に、話し言葉・書き言葉変換の基本的な手法を組み込んだ自然言語処理システムを試作した。このシステムでは、話し言葉を、違和感の生じる表現、音声に適さない表現、複雑な構造をもつ表現を含まない、「音声として聞いたときにわかりやすい表現」と定義し、機能語的な表現を中心に、普通体から丁寧体への文法的な変換を行う。

第四に、自然言語テキストからプレゼンテーション補助資料を生成する手法の開発を行った。この手法では、数文から数十文の意味的まとまりを持つ自然言語テキストを入力とし、文章構造の解析、主題や重要な説明表現の抽出、見出し語・重要説明表現の配置を順次行うことによって、要約スライドを作成する。

第五に、比ゆ的な意味の可視化と、重要語のスーパーインポーズによって、言語モダリティと非言語モダリティの同期を強調する手法を開発した。与られた自然言語テキストは、まず日本語解析エンジンを用いて解析され、非言語的なモダリティによって強調すべき部分が同定され、その部分をどのような非言語モダリティによって強調するのかが決定される。例えば、重要な概念について述べるときには、強調のジェスチャーを用い、眉を上げて目を大きく見開くというプランが用いられる。次に、自然言語テキストが音声合成システムに入力され、音声ナレーションが作成される。同時に、音声とエージェントアニメーションを同期させるためのタイムスケジュールが計算される。

2.3 非言語短発声の韻律的特徴の語用的観点からの役割分析（Ward）[4]

uh-huh など、言語的な内容の伝達を主目的としない短い発話の韻律は、対話、会話において話者の心理状態、意図などの伝達に重要な役割を果たす。このような音声の韻律的特徴を分析し、発話内容とは関係なく、以下のように特定の情報の伝達に寄与することを明らかにした。

1. 音節への区分：発話忌避
2. 長さ：思考の程度
3. 強さ：確信・重要性
4. ピッチの高さ：興味の程度

5. ピッチの傾き：理解の程度
6. 音質（きしみ声）：根拠に基づいた確信（専門知識に基づく意見の発言を示す等）。

興味あることに、これらの韻律的特徴は通常の文で見られる傾向とも一致している。例外は、声帯振動の様子が通常と異なるとき、笑いながらしゃべっているとき、繰り返すとき、などで見られた。このような発話の対話システムへの利用を進めた。

2.4 生命的エージェントによる感性的マルチモーダルコンテンツ記述と生成（石塚）[5]

顔と姿を持ち音声機能を有する生命的エージェント（lifelike agents；擬人化エージェントやECA (embodied conversational agents) などとも呼ばれる）を用いるマルチモーダルインタフェースやコンテンツが出現し始め、複雑化が進行する情報化社会の中で今後も理解しやすく親しみやすい新形態のマルチモーダルメディアとして発展が期待されている。高度音声合成の応用として、この生命的エージェントに関しコンテンツ記述言語MPMLと感性的機能を中心に研究開発を実施し、実用性のある成果を得ている。

MPMLはマークアップタグによるXML言語であり、VB ScriptやJava Scriptプログラムのようにプログラミング言語を知らなくても、HTMLを記述できる人なら新たな20〜30程のタグの使い方を知ることにより記述可能である。（HTMLのようにMPMLのGraphical Editorの初期版も用意されているので，MPML自体をたとえ知らなくても記述することは可能になっている。）ビデオや音声データも含むメディア同期用にSMILの基本機能を含んでいる。キャラクタエージェントとしてはMS Agentsを基本的にサポートしているが、ドライバ部分のプログラムを書くことにより各種エージェントを使うことができ、実際、3D VRML空間でのH-Anim規格のエージェント、携帯電話 (i-mode, au, J-phone) 上のエージェントにも対応するようになっている。顔表情豊かな独自作成のSmArt Agentにも対応している。

エージェントのプレゼンテーションは平板なものになりがちだが、感情表現の付加はエージェントの生命感、信頼感を向上させる上で重要である。エージェントの感情は視聴者の感情も呼び起こし、親近感、エンタテイメント性、モチベーション等を向上させる効果を有する。感情は喜び（幸福感）、悲しみ、驚き、怒り、嫌悪、恐れ・・などの言葉で語られるが、場合によりそのカテゴリ分けは不統一で、根拠も不十分なものであった。MPMLでは最も包括的なOCCモデルによる感情を扱うようにしており、感情状態をタグで囲んで記述する。この時、発話の前後に感情による動作をし，感情に応じてTTSの発話スピード、ピッチ、ピッチの変動幅，強度の音声パラメータを変化させるようにしている。これにより少ない記述で、音声を含めて感性的なエージェントを生成できるようにしている。

2.5 多様な話者性および発話スタイル・感情表現による音声合成のための韻律生成（小林）[6]

多様な話者性による音声合成の実現をめざし、そのための統計的モデル（HMM）に基づいた韻律生成手法および音声合成システムの開発を行った。具体的には、多様な話者性による音声合成を実現するために必要となる任意話者の声質・韻律の生成と様々な発話スタイルや感情を表現するための韻律生成について検討した。

まず、任意話者の声質および韻律特徴による音声の合成手法に対し、かねてより提案してきた平均声モデルと話者適応に基づく手法について種々の検討を行った。平均声モデルとは、HMMに基づく音声合成において、複数話者の音声データベースから学習された音声単位HMMのことであり、これを用いて合成された音声は複数話者の平均的な声質および韻律特徴を持つと考えられることから、これを平均声と呼んでいる。任意話者の音声を合成するには、対象となる話者の少量の発声データを用いて平均声モデルを話者適応技術によりモデル適応した後、HMM音声合成に基づいて韻律およびスペクトルパラメータの生成を行う。本研究では、モデル適応手法として、最尤線形回帰（MLLR）に基づいたスペクトルおよび韻律モデルの適応手法を開発すると同時に、平均声モデルを学習する際、各話者の音声データが大量に存在しない場合にも合成音声の自然性を劣化させないモデル学習法として共有決定木コンテキストクラスタリング（STC）手法を、さらに合成音声の品質を向上させる話者適応学習を組込んだ学習法を提案し、その有効性を示した。

一方、多様な発話スタイルや感情を表現するための韻律生成手法の開発では、まず男女各１名が「丁寧」「ぞんざい」「楽しげ」「悲嘆」の４種類のスタイルにより読み上げた503文章からなるスタイル音声データ

ベースを作成した。そして、これを用いてHMM音声合成のための二つのスタイルモデルリング手法を提案し、その評価を行った。さらに、多様なスタイルを実現する韻律・スペクトル生成手法として、スタイル補間手法ならびにスタイル適応手法を提案した。スタイル補間手法では、異なるスタイルに対応する音声合成単位 HMM の出力分布をモデル補間することにより、その中間的なスタイルによる音声を合成可能なこと、また補間比率を徐々に変化させることにより、あるスタイルから他のスタイルに滑らかに変化する音声を合成できるスタイルモーフィング技術を開発した。スタイル適応手法では、あるスタイルによる少量の発声データが与えられた際に、読み上げ調のモデルからそのスタイルにモデル適応することにより、そのスタイルで任意のテキストに対応する音声が合成できることを示した。

この他、瞬時周波数振幅スペクトルの調波構造を利用し、F0抽出を高精度に行う手法を確立した。信頼性の高い統計的モデルを自動構築するには、大量の F0 データを、種々の音声データベースから自動的に精度良く抽出することが求められる。これに対して、瞬時周波数振幅スペクトルの調波構造を利用した F0抽出法を開発した。この方法では、F0を求める際に用いる周波数帯域においてどれだけ明確な調波構造を成しているかを示す尺度となる調波構造指数を定義し、これに基づいて適切な周波数帯域及び分析窓長を自動選択するものであり、有声区間から無声／無声区間への切り替り部分での抽出精度を従来法に比べ向上させることが可能であることを明らかにした。

2.6 コーパスに基づいた統計的なアプローチによる韻律のモデル化・生成手法の確立（徳田）[7]

HMM に基づいた一連の手法を提案し、感情音声他への適用を行い、多様な音声合成が可能となることを示した。以下に、具体的な成果をまとめる。

韻律の統計的モデル化手法の確立：無声区間を含む（F0の無い区間）を含む音声の基本周波数パターンを、直接、HMM によりモデル化するため、多空間分布 HMM (MSD-HMM) と呼ぶ新たなHMMを定義し、拡張されたHMMのモデルパラメータ推定手法を与えた。これらにより、音声の基本周波数パターンを理論的整合性をもって統計的にモデル化することが可能となった。また、学習により得られたHMMから、基本周波数パターンを生成し、自然性の高い基本周波数パターンが生成できていることを確認した。

HMMによる韻律生成手法の確立：これまでに開発したHMMから基本周波数パターンを生成する手法を更に高性能化するため、「ガンマ分布による継続時間長モデル」を導入した。これにより、より少ないHMMのモデルパラメータ数で同等の基本周波数パターンを生成できることを示した。また、「有声・無声境界でのダイナミックスを考慮した基本周波数パターンのモデル化手法」を導入し、生成される基本周波数パターンの自然性の向上をはかった。

感情を含む韻律の統計的モデル化手法の開発：これまでの成果をもとに、多様な声質、感情、発話スタイルの韻律生成が可能なシステムを開発し、合成音声の評価を行った。特に、音声データベースの韻律ラベルの自動付与、感情音声データベースのラベリング手法などについても検討を行った。また、発話スタイル、感情のモデル化を考慮したモデル学習アルゴリズムについて更なる改良を行った。

統計的モデルを用いた韻律生成システムの開発：以上の成果をもとに、多様な声質、感情・発話スタイルの韻律の生成が可能なシステムを開発し、音声合成によって評価した。また、「固有声手法を」導入することにより、ユーザが好みの声質を自在に設定可能なシステムを構築した。更に、システムを多言語に拡張し、日本語のみならず、英語、中国語、ポルトガル語などでの動作を確認した。

3. 今後の発展

本研究では、生成過程モデルに基づく感情音声のコーパスベース韻律制御、語用論的立場からの非言語発話の韻律的特徴の分析、感情の統合的記述とマルチ・モーダル・エージェントにおける感情表現、平均音声モデルと適応手法を用いたHMM音声合成による種々の声質・発話スタイルの合成音声、統合的なHMM合成手法の確立と感情音声合成、等の成果を達成した。これらは、学会誌論文、国際会議論文等として発表する他、HMM 音声合成ソフトウエアは共用ソフトウエアとして公開している。また、コーパスベース音声合成に必要な生成過程モデルに基づく韻律コーパスも公開している。

音声合成の最終的な目標は、その利用者が、希望す

る声質・調子の音声を自由に合成し得るということであろう。このためには声質・調子を対応した適切な韻律制御が重要である。本研究はこの観点から、合成すべき声質・調子の音声コーパスを用意し、それを、回帰木やHMMなどの統計的手法により処理して、韻律制御の方法を自動的に学習することを、1つの大きな柱として研究を進めた。ただし、合成すべき声質・調子の十分な量の音声コーパスが得られるとは限らず、そのための対処として、少数のコーパスからHMMの適応手法によって、必要とする声質・調子の音声を合成する手法の開発を行った。今後は、HMM合成で問題となっている音質の向上について研究を進めるとともに、コーパスが得られないような声質・調子の音声を、内挿、外挿等により実現するための方策を開発することを進める。この場合の問題点として、韻律の構造は、話者、発話速度、スタイルによって必ずしも同じでないと言うことである。コーパスベース手法のみならず、従来のルールベース手法の考え方を取り入れた手法の開発が重要となる。

一般に、音声合成というと、テキストを入力として音声を出力とするテキスト音声合成が想起されるが、機械から人間への情報伝達を考えた場合、伝達する情報を文章化してさらに音声化するプロセスが必要になる。これは、概念音声合成と呼ばれるが、文章化に際し、テキスト解析では得られない高次の言語情報が得られ、これを適切に反映した音声合成が求められる。この観点から、本研究での成果も利用し、場面に適合した、情報から音声への一貫した変換手法を可能とする手法の開発を進める。

参考文献

[1] Keikichi Hirose, Kentaro Sato, and Nobuaki Minematsu, "Corpus-based synthesis of fundamental frequency contours with various speaking styles from text using F0 contour generation process model" Proc. 5th ISCA Speech Synthesis Workshop, Pittsburgh, pp.162-166 (2004-6).

[2] Shuichi Narusawa, Nobuaki Minematsu, Keikichi Hirose and Hiroya Fujiaski, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," Proc. IEEE International Conference on Acoustics, Speech, & Signal Processing, Orlando, Vol.I, pp.509-512 (2002-5).

[3] Q. Li, Y. Nakano, M. Okamoto, and T. Nishida, "Highlighting multimodal synchronization for embodied conversational agent", Proc. 2nd International Conference on Information Technology for Application (ICITA2004), pp.50-55 (2004-1).

[4] Nigel Ward, "Pragmatic functions of prosodic features in non-lexical utterances", Proc. International Conference on Speech Prosody, pp 325-328 (2004-3).

[5] Helmut Prendinger, Sylvain Descamps, and Mitsuru Ishizuka, "MPML: A markup language for controlling the behavior of life-like characters," Journal of Visual Languages and Computing, Vol.15, No.2, pp.183-203 (2004-4).

[6] Junichi Yamagishi, Koji Onishi, Takashi Masuko, and Takao Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH '03), Geneva Vol.III, pp.2461-2464 (2003-9).

[7] Keiichi Tokuda, Takashi Mausko, Noboru Miyazaki, Takao Kobayashi, "Multi-space probability distribution HMM," IEICE Trans. Information and Systems, Vol.E85-D, No.3, pp.455-464 (2002-3).