社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－SLP－50　(8)
2004／2／7

生成過程モデルに基づく コーパスベース感情音声合成とその評価

佐藤 賢太郎[1],　広瀬 啓吉[1],　峯松 信明[2]

[1] 東京大学大学院 新領域創成科学研究科,　[2] 東京大学大学院 情報理工学系研究科
Tel.: 03–5841–6393,　Fax.: 03–5841–6648
{kentaro, hirose, mine}@gavo.t.u-tokyo.ac.jp

あらまし　基本周波数生成過程モデルの制約下で、感情音声の基本周波数パターンを推定するコーパスベース手法の開発を従来から行っている。この手法では、モデルの指令の推定を行う統計的手法として二分回帰木を用いている。ここでは、この手法を高精度化すると共に、同様の手法によりアクセント句を推定することを行い、漢字仮名混じりテキストを入力として、感情音声を合成するシステムを構築した。システムを用いて合成した「怒り」「喜び」「悲しみ」の各感情音声を用いて、若干名の日本人話者に評価をさせたところ、「怒り」については (推定の目標とする F_0 パターンを用いた場合と近いという) 高い評価を得た。

キーワード　感情音声、コーパス、統計的手法、基本周波数パターン、生成過程モデル

Corpus-Based Emotional Speech Synthesis Based on Generation Process Model and Its Evaluation

Kentaro Sato[1],　Keikichi Hirose[1]　and　Nobuaki Minematsu[2]

[1] Graduate School of Frontier Sciences, University of Tokyo,
[2] Graduate School of Information Science and Technology, University of Tokyo,
Tel.: 03–5841–6393,　Fax.: 03–5841–6648
{kentaro, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract　We have been developing a corpus-based method for generating F_0 contours of emotional speech under the constraint of the functional model of F_0 contour generation process. The Method uses the binary regression tree as the statistic method to estimate the model commands. In the current paper, the method was improved from several aspects and a similar method was added for the estimation of accent phrase boundaries in order to develop a system for synthesizing emotional speech from a text input. Using the system, we synthesized “angry,” “joyful,” and “sad” speech from the text input. The speech was evaluated by Japanese listeners, and obtained a rather good score for “anger” (a score close to the case of using target F_0 contour).

Key words　Emotional Speech, Corpus, Statistical Method, Fundamental Frequency Contours, Generation Process Model

1. はじめに

近年、マルチメディア技術の飛躍的な向上などによって、多種多様な情報通信サービスが整備されてきている。その中で、人間とのインタフェースの１つとして期待される音声についても、機械的でない、人間味を帯びたコミュニケーションを実現する研究が盛んに行われている。

文字情報では、文脈などによって言語情報以外の情報が伝達されるが、音声言語では、文字言語に比べ、態度や感情といった情報の比重が増す。これらの情報は主に、韻律的特徴によって伝えられることから、我々は韻律の制御法を中心に研究を進めている。本稿では、様々な発話スタイルの中から、感情音声に焦点を当てた。

既に、感情音声合成についての研究は数多く行われている。飯田ら [1] は波形接続方式による感情音声合成を提案している。また、朗読音声の合成において、現在盛んに研究されている HMM の枠組みを感情音声に適用し、感情音声合成を実現しようとする試みもある [2][3]。しかし、韻律の制御については、未だ不十分な部分が多く、感情音声については特に、その制御法の開発が重要な課題である。

その中で、我々は、基本周波数パターン生成過程モデル (以下、F_0 モデル) [4] と統計モデルを用いての韻律制御を提案し、実用化を目指している。既に、感情音声についての F_0 モデルパラメータや音素持続時間長を統計的手法によって推定することを行い、同手法を用いた朗読音声の推定に近い結果を得ている [5]。

本稿では、テキストから感情音声を合成するシステムを構築するために、従来の手法 [5] を改善した点、さらに、アクセント句の推定など、追加した枠組みについて述べる。また、本手法を「平静」「怒り」「喜び」「悲しみ」のそれぞれの感情をこめて読み上げた音声に適用し、得られた音声について考察した。

2. 使用するモデル

2.1 基本周波数パターン生成過程モデル

本稿では、抽出した基本周波数パターンを分析するにあたって、下記の基本周波数パターン生成過程モデルを用いている [4]。本モデルは、少ないパラメータで F_0 パターンを良く近似するため、TTS システムに用いる上で利点が多い。

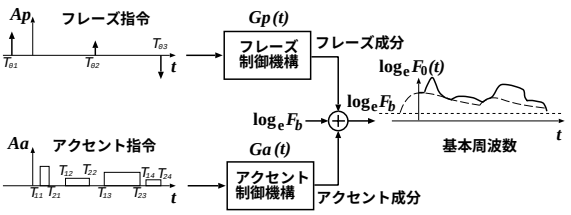

図 1. F_0 パターン生成過程モデル

図 1 に示すように、このモデルでは、比較的ゆっくりとした土台の起伏部分（フレーズ成分）と、比較的急速に上下する成分（アクセント成分）とに分けて考えている。式で表わすと、対数基本周波数の時間パターン $\ln\ F_0(\mathrm{t})$ は、

$$\begin{aligned}\log_e F_0(t) = \log_e F_b + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) \\ + \sum_{j=1}^{J} A_{aj}\{G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j})\}\end{aligned} \tag{1}$$

と表される。右辺第 2 項がフレーズ成分、第 3 項がアクセント成分にあたり、A_{pi} と T_{0i} はそれぞれ i 番目のフレーズ指令（インパルス）の大きさと生起位置、A_{aj} と T_{1j} と T_{2j} はそれぞれ j 番目のアクセント指令（ステップ）の振幅と立上り位置、立下り位置である。また、G_{pi}、G_{aj} は

$$G_{pi}(t) = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t), & t \geq 0, \\ 0, & t < 0, \end{cases} \tag{2}$$

$$G_{aj}(t) = \begin{cases} \min[1 - (1 + \beta_j t)\exp(-\beta_j t), \gamma], & t \geq 0, \\ 0, & t < 0. \end{cases} \tag{3}$$

と近似される。ここで α、β はそれぞれの制御機構の固有角周波数、γ はアクセント成分が有限時間内に一定値に達することを保証する相対飽和値である。α、β および γ の話者ごと、発話ごとの変動は比較的小さいため、初期値としては、それぞれ $\alpha = 3.0\mathrm{rad/s}$、$\beta = 20.0\mathrm{rad/s}$、$\gamma = 0.9$ を用いることができ [4]、本稿ではこの値に固定してモデル化している。

2.2 統計モデル

韻律制御に用いられる統計的手法には、重回帰分析・ニューラルネットワークなど様々なものが

あるが、今回の実験では、決定木を使用した。決定木を用いた韻律生成は、他の統計的手法と比べても、同程度の結果を得ており [6]、また、構築されたモデルに関する解析が容易であるという利点を持つ。

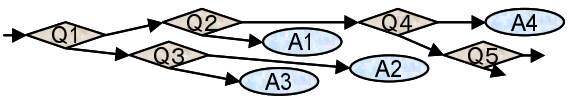

図 2. 決定木の概略図

決定木は、図 2 に概略を示すような統計モデルであり、各ノードに用意されている質問に答えて階層を進んでいくと、最終的に必ずひとつの答えにたどり着き、所望の値が得られるというものである。また最終的得られる答えが、離散的な値あるいはクラスといったようなものの場合のモデルを決定木、連続値を答えに持つような木を回帰木と呼ぶ。本稿ではこれら 2 つをまとめて決定木と呼ぶことにする。図 2 に示したのは、2 分木形態の場合の決定木である。

決定木は要因の組み合わせによってデータを逐次的に分割してゆくことによって作成される。学習は各ノード作成時点で最も効果的な要因による分割が選ばれるという最良探索法によるため、要因の組み合わせすべてによって作成可能な木からの全数探索はなされていない。結果として得られる木の全体的な最適性は必ずしも保証されないが、データ中に見られる要因による分布の偏りをもとに統計的に有為な範囲でモデルが得られるためデータ量に応じたモデル化が可能である。また、各要因による効果を線形回帰モデルのように一定値で表すことをしないため、要因間にまたがった効果を表現する上で自由度が高い。

決定木構築手法としては、CART(Classification And Regression Trees) [7] によるモデル化が挙げられる。今回用いる決定木もこの CART の手法を用いた The Edinburgh Speech Tools Library [8] の wagon を採用している。

決定木構築の際のリーフ数は、細かくすればするほど、学習データを良く再現するが、未学習データについての推定精度が悪くなってしまう。本稿では、最小リーフ数を 40 に設定して実験を行っている。

3. 韻律データベース

3.1 使用した音声資料

使用した音声試料は、女性話者 1 名により発話されたもので、平静音声は ATR 連続音声データベースの 503 文を読み上げたもの、感情音声は各感情ごとに用意されたそれぞれ数百文を各感情をこめて読んだものである。使用した各文はいずれも各感情をこめて発話し易いような文である。音声は全てサンプリング周波数 10KHz、16bit 直線量子化したものである。

3.2 韻律データベースの自動作成

統計モデルの学習に用いるデータは、大量であるため、それらの構築を自動で行った。以下に具体的な流れを示す。

1. 音声ファイルから基本周波数値を抽出し、F_0 モデルパラメータ自動分析システム [9] により F_0 モデルパラメータファイルを得る。ここで、各感情の基底周波数 F_b は、平均値から標準偏差の 3 倍を引いたものに固定している。各感情での F_b は、表 1 の通りである。

表 1. 各感情における F_b

	平静	怒り	喜び	悲しみ
F_b (Hz)	147.67	182.49	210.30	182.49

2. 音声ファイルと、漢字仮名交じり文から得た発音ファイルから Julius[10] を利用して、音素セグメンテーションファイルを得る。なお、漢字仮名交じり文から発音ファイルを得るには、Chasen[11] の機能を利用した。

3. 漢字仮名交じり文から、Chasen を用いて形態素、品詞情報を得るとともに、JUMAN+KNP[12] を用いて文節、統語情報を得て、言語情報ファイルとする。

4. 音素セグメンテーションファイルと言語情報ファイルを参照して、実際の音声の時間情報と個々の形態素のマッチングを取る。

5. F_0 モデルパラメータファイルを参照し、先行アクセント指令の立ち下がり位置と当該アク

セント指令の立ち上がり位置の間にある形態素境界をアクセント句境界とする。ただし、形態素境界が複数ある場合は、当該アクセント指令の立ち上がり位置に一番近い文節境界（文節境界がない場合は、形態素境界）をアクセント句境界とする。なお、文頭と文末には必ずアクセント句境界があるものとしている。

6. アクセント結合規則 [15] に基づき、各アクセント句のアクセント型を決定する。アクセント核をなす各モーラ母音開始時点とアクセント指令の立ち下がり位置との差分を T_{2off} として定義する。また、アクセント指令の生起タイミング T_{1off} をアクセント句の先頭モーラの母音開始時点からの差分とする。

7. フレーズ指令はアクセント指令の間に 0 または 1 個存在すると仮定して検索を行い、存在した場合、時間的に後続するアクセント句の先頭モーラの母音開始時点からの差分をフレーズ指令のタイミング T_{0off} とする。

8. 以上の作業を、用意したファイルすべてについて行い、韻律データベースを作成する。

各プロセスにおいて、抽出エラーを含むものは学習データから除いている。実際に用いたデータ数を表 2 に示す。なお、close は学習に用いたデータ、open は学習に用いなかったテスト用のデータを表わす。

表 2. 用いた音声試料数

	平静		怒り		喜び		悲しみ	
	close	open	close	open	close	open	close	open
文数	333	50	472	50	358	50	305	50
形態素数	5050	734	7770	852	5219	558	4880	809
アクセント句数	2340	338	3247	346	2391	271	2185	389

4. アクセント句推定実験

4.1 アクセント句推定の入出力項目

F_0 モデルパラメータの学習・推定の単位はアクセント句である。アクセント句とは、アクセント成分を一つ含む日本語の発声単位である。アクセント句は文字言語上ではおおむね、文節として表わされる。しかし、アクセント句は、音声言語特有の韻律現象であるため、話者ごと、あるいは同じ話者でも発話ごと、感情ごとに異なる場合がある。

TTS システムを考える際には、言語情報からアクセント句を推定しなくてはならない。本実験ではアクセント句を統計モデルから推定するという方法をとった。構築に用いる言語情報の単位は形態素である。表 3 にアクセント句推定モデルの入力項目を示す。

表 3. アクセント句推定の入力項目

形態素の情報	カテゴリ数
当該 (先行) 形態素の品詞	15(16)
当該 (先行) 形態素の活用形	24(25)
当該 (先行) 形態素の活用型	35(36)
当該 (先行) 形態素のモーラ数	10(11)
当該形態素の文内位置	57
当該形態素の属する文節の境界コード	21
当該形態素の先頭の文節境界の有無	2 値

カテゴリ数とは、用いた韻律データベース中に出現した該当要因のとりえた範囲内の値の種類数である。なお、出力項目はアクセント句境界の有無を表わすフラグ（2 値）である。

4.2 推定結果と考察

推定結果を表 4 に示す。どの感情でも 80 %前後の正解率を得ていて、選択した入力項目で良好な推定が行われていると言える。正解データは、音声より自動抽出 [9] された F_0 モデルパラメータとしているが、自動抽出の際に言語情報は考慮されておらず、人間の発話スタイルと一致したパラメータ抽出がなされていないパターンも多数存在すると考えられる。それらが、誤り率を増加させる 1 つの原因として考えられる。

表 4. アクセント句推定結果 (%)

	平静		怒り		喜び		悲しみ	
	close	open	close	open	close	open	close	open
正解率	86.2	84.3	83.6	82.0	80.0	76.7	81.0	79.5
挿入誤り率	9.3	11.0	10.2	11.0	12.4	14.1	10.7	12.0
脱落誤り率	4.4	4.6	6.2	6.9	7.6	9.1	8.3	8.5

5. F_0 モデルパラメータ推定実験

5.1 F_0 モデルパラメータ推定の出力項目

テキストから F_0 パターンを作成するために、推定する必要があるパラメータは表 5 の通りである。

PF と A_p と T_{0off} はフレーズ指令に起因する F_0 モデルパラメータである。PF は当該アクセント句の先頭にフレーズ指令が存在するかどうかのフラグである。例えば、PF が 1 であれば、T_{0off} の位置に A_p の大きさでフレーズ指令が立つものとす

る。逆に PF が0であった場合は、A_p と T_{0off} の値はそのアクセント句内で0となる。A_a と T_{1off} と T_{2off} はアクセント指令に起因するパラメータである。アクセント指令はアクセント句内に必ず一つ存在するもので、その生起・終了位置が T_{1off}・T_{2off} であり、A_a はその大きさを表している。

表 5. F_0 モデルパラメータ推定の出力項目

出力項目	カテゴリ数
先頭のフレーズ指令有無 PF	2 値
フレーズ指令の大きさ A_p	連続値
フレーズ指令のタイミング T_{0off}	連続値
アクセント指令の大きさ A_a	連続値
アクセント指令の生起タイミング T_{1off}	連続値
アクセント指令の終了タイミング T_{2off}	連続値

5.2 F_0 モデルパラメータ推定の入力項目

統計モデルでは、計算によって自身の説明に有力な入力要因を取り込んで自動的に学習するが、そのためには、あらかじめ必要とされる要因を準備し、モデル化されやすいように適切にコード化しておくことが重要である。また、用いる情報は全て漢字仮名交じり文から自動的に得られるものである必要がある。これらをふまえた上で入力項目としては、表 6 のようなものを与えている。学習の単位はアクセント句である。

表 6. F_0 モデルパラメータ推定の入力項目

出力項目	カテゴリ数
当該句の文内位置	27
当該 (先行句) の有するモーラ数	28(29)
当該 (先行句) のアクセント型	19(20)
当該 (先行句) の有する単語数	11(12)
当該 (先行句) の最初の単語の品詞	14(15)
当該 (先行句) の最初の単語の活用形	21(22)
当該 (先行句) の最後の単語の品詞	14(15)
当該 (先行句) の最後の単語の活用形	21(22)
先頭の境界コード	18
当該句の PF	2 値 *
当該句の A_p	連続値 *
当該句の T_{0off}	連続値 *

*:二段階推定（A_a,T_{1off},T_{2off} のみ）

境界コードは、KNP の出力から計算されるもので、文節間の係受け情報とその深さを表わす。なお、A_a・T_{1off}・T_{2off} については、言語情報から推定された PF・A_p・T_{0off} を新たに入力項目として加えることによって、推定精度を上げることを試みている。

5.3 推定結果と考察

2 つの F_0 パターンの違いを定量的に表す尺度として式 4 に示す F_0MSE を用いる。ただし、t は有声フレーム数、T は有声フレーム総数である。

$$F_0MSE = \frac{\sum_t (\Delta \ln F_0(t))^2}{T} \quad (4)$$

表 6 の項目を入力とし、F_0 モデルパラメータ推定の決定木を構築、評価を行った。表 7 に各アクセント句におけるフレーズ指令挿入の正解率、表 8 に F_0MSE を示す。

表 7. PF 正解率（%）

	close	open
平静	78.6	74.4
怒り	74.7	70.8
喜び	72.5	69.4
悲しみ	78.7	74.6

表 8. F_0MSE 平均値

	close	open
平静	0.045	0.049
怒り	0.040	0.056
喜び	0.039	0.052
悲しみ	0.031	0.033

F_0MSE について、先行研究 [5] に比べ低下が見られた。これは、アクセント核をアクセント結合規則 [15] に基いて求めるよう改善を施したからである。また、表 7 を見ると、フレーズ指令の推定誤りがかなりあることが分かる。このフレーズ指令の推定誤りが誤差の大きな原因と考えられる。

6. 音素持続長時間推定

音素持続時間長の推定も、アクセント句や F_0 モデルパラメータの推定と同様の枠組みで行っている。今回の実験では、先行研究 [5] の手法を用い推定を行った。

7. 感情音声の合成と評価

7.1 音声合成の条件

怒り・喜び・悲しみの各感情について、以上の統計モデルを元に、音声を合成し、聴取実験を行った。

スペクトルについては、HMM 音声合成ツールキット [13] を用いて作成した。学習用データにはアクセント句、F_0 モデルパラメータ学習の際と同じデータを用い、怒り・喜び・悲しみの各感情についてモデルを作成した。

サンプリング周期 16kHz、フレーム周期 5ms で、長さの 25ms の Hamming 窓を用い、0〜24 次のメルケプスラム、Δ および Δ^2 メルケプストラムの計 75 次元の特徴ベクトルを作成した。なお、HMM は left-to-right トライフォンモデルで状態数は 7 である。また、合成には音声信号処理ツールキット SPTK[14] の MLSA フィルタを用いた。

評価用の文としては、アクセント句の推定エラー、フレーズ指令の推定エラーのなかった文のうち、各感情で 10 文ずつを任意に選択した。同様

に、自動抽出した F_0 パターン、音素持続時間長を付与した文を各感情でテストした。被験者は日本語話者 15 名である。

7.2 評価条件と分析結果

各文について、平静・怒り・喜び・悲しみの 4 感情のうち、どの感情に聞こえるか判別してもらった。その結果を表 9 に示す。なお、「正解」は HMM でのスペクトル生成に自動抽出した F_0 パターン・音素持続時間長を付与した文、「推定」は推定した F_0 パターン・音素持続時間長を付与した文である。

表 9. 感情の判別率 (%)

	怒り		喜び		悲しみ	
	正解	推定	正解	推定	正解	推定
平静	2.2	5.2	0	47.3	8.9	30.9
怒り	**93.3***	**87.4***	2.2	9.3	2.2	6.0
喜び	2.2	2.2	**97.8***	**38.7***	2.2	6.7
悲しみ	2.2	5.2	0	4.7	**86.7***	**56.4***

*：正解を判別した確率 (怒り・喜び・悲しみ)

また、同時に、各文についてどの程度の感情が含まれているか 1〜5 の 5 段階 (5 が最も感情が含まれている) で評価してもらった。その平均値を表 10 に示す。なお、感情の判別が間違っているものについては 0 とした。

表 10. 感情の大きさ (5 段階評価)

	怒り		喜び		悲しみ	
	正解	推定	正解	推定	正解	推定
評価値	3.09	3.08	3.38	1.03	3.64	1.34

7.3 考察

怒りについては、推定した韻律を付与した文でも高い判別率を得ている。原因として、今回用いた話者では、特に怒りで、F_0 のダイナミックレンジが大きくなる、音素持続時間長が短くなるなどの顕著な特徴が見られたということが考えられる。

また、各感情において、平静と誤判別される割合が大きい。本稿でアクセント型は、アクセント句推定から得られる文にアクセント結合規則 [15] を適用することで得ているが、この規則は平静音声について求めたもので、感情音声にそのまま適用することの可否についてはさらに検討が必要である。

表 10 に示す感情の大きさを見てみると、怒りについては、正解の韻律に近い結果を得ているが、喜び・悲しみでは正解とかなりの差が見られた。これらは、判別率の結果に起因するところが大きい。しかし、怒りのように、F_0 のダイナミックレンジの大きなものについては、対応できているが、小さなものについては、まだ推定精度が低いというのも大きな原因である。

8. まとめ

テキストからの音声合成において、言語情報を入力とした韻律生成を行うことによって感情音声を実現しようとする枠組を実装した。聴取実験において、怒りについては正解の韻律に近い評価を得た。

参考文献

[1] A.Iida, F.Higuchi, N.Campbell, M.Yasumura : "Corpus-based speech synthesis system with emotion," *Speech Communication*, Vol.40/1-2, pp.161-187 (2002).

[2] J.Yamagishi, K.Onishi, T.Masuko, T.Kobayashi : "Modeling of Various Speaking Styles and Emotions for HMM-Based Speech Synthesis," *Proc. EUROSPEECH*, Vol.4, pp.2461-2464 (2003).

[3] 都築亮介, 全炳河, 徳田恵一, 北村正, Murtaza Bulut, Shrikanth S.Narayanan : "HMM に基づく感情音声合成に関する検討,", 日本音響学会秋季講演論文集, pp.241-242 (2003.9).

[4] H.Fujisaki, S.Nagashima : "A model for synthesis of pitch contours of connected speech," *Annual Report of Engineering Research Institute, University of Tokyo*, vol.28, pp.53-60 (1969).

[5] 桂聡哉, 広瀬啓吉, 峯松信明 : "感情音声のための生成過程モデルに基づくコーパスベース韻律生成とその評価," 電子情報通信学会技術研究報告, SP2002-184 (2003.3).

[6] 江藤雅哉, 広瀬啓吉, 峯松信明 : "テキスト音声合成システムのための統計モデルによる F_0 パターン生成の改良," 日本音響学会春季講演論文集, pp.245-246 (2002.3).

[7] L.Brieman, J.H.Friedman, R.A.Olshen, and C.J.Stone : "Classification and Regression Trees," *Wadsworth, Paciic Grove, California* (1984).

[8] The Edinburgh Speech Tools Library. http://www.cstr.ed.ac.uk/projects/speech_tools/

[9] 成澤修一, 峯松信明, 広瀬啓吉, 藤崎博也 : "声の基本周波数パターン生成過程モデルのパラメータ自動抽出法", 情報処理学会論文誌, Vol.43, No.7, pp2155-2168 (2002).

[10] 大語彙連続音声認識デコーダ Julius. http://winnie.kuis.kyoto-u.ac.jp/pub/julius/ .

[11] 形態素解析システム 茶筌. http://chasen.aist-nara.ac.jp .

[12] 日本語構文解析システム KNP. http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/ .

[13] HMM 音声合成ツールキット HTS. http://hts.ics.nitech.ac.jp/

[14] Speech Signal Processing Toolkit. http://kt-lab.ics.nitech.ac.jp/ tokuda/SPTK/ .

[15] 喜多竜二, 峯松信明, 広瀬啓吉 : "日本語テキスト音声合成を目的としたアクセント結合規則の構築と改良", 電子情報通信学会技術研究報告, SP2002-26 (2002.5).