生成過程モデルに基づくコーパスベース基本周波数パターン生成[#]

Corpus-based generation of fundamental frequency contours based on the generation process model

東京大学大学院情報理工学系研究科

Graduate School of Information Science and Technology, University of Tokyo

広瀬　啓吉

Keikichi Hirose

＜研究協力者＞

東京大学大学院新領域創成科学研究科

Graduate School of Frontier Informatics, University of Tokyo

峯松　信明

Nobuaki Minematsu

東京大学大学院新領域創成科学研究科

Graduate School of Frontier Informatics, University of Tokyo
*Currently with NEC

佐藤賢太郎*

Kentaro Sato

A corpus-based method of generating fundamental frequency (F_0) contours from text was developed. Instead of directly predicting F_0 values, the method predicts command values of the F_0 contour generation process model using binary decision tree. Because of the model constraint, synthetic speech with certain quality is available even when the prediction is done incorrectly. Also, it is easy to incorporate our knowledge on the F_0 contours in the method by manually adjusting command values after the prediction. The method includes a scheme of extracting the model commands from given F_0 contours, which makes it possible to prepare the training corpora automatically. F_0 contours of any speaking styles can be generated if the training corpora in such styles are available. As an example of various styles, emotional speech (anger, joy, sadness) was selected as the research target as well as the calm speech. Since accuracy in the extraction of the model command values is crucial for the method, a constraint is applied on the position of phrase commands. Also, since performance of phrase command prediction dominates the overall accuracy of generated F_0 contours, the method was modified from its original version to predict phrase commands first. The mismatches between the predicted and target contours for angry speech were similar to those for calm speech. Synthesis of emotional speech was conducted with text inputs. The segmental features were handled by the HMM-based speech synthesis method and the phoneme durations are predicted in a similar corpus-based method. Perceptual experiment was conducted for the synthesized speech, and the result indicated that the anger could be well conveyed by the developed method. The result came worse for joy and sadness.

Key Words: Fundamental frequency (F_0) contour, Generation process model, Corpus-based method, HMM-based speech synthesis, Emotional speech

[#] 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

1. 研究の目的と概略

波形編集方式に基づくコーパスベース音声合成により、合成音声の品質向上には目覚しいものがあるが、それによって、あらためて韻律の不自然さという問題点がクローズアップされる結果となっている。さらに、従来の音声合成で生成される音声は、ほぼ朗読調（アナウンス調）に限られていたが、対話システムのユーザにとって受け入れやすい音声出力、アミューズメント用途の合成音声など、言語情報のみならず、意図、感情、個性と言ったパラ言語・非言語情報を的確に反映した種々の調子の音声を合成することに対するニーズが増大している。本研究は、このような音声合成を可能とする技術としての、韻律制御手法を開発することを目的としている。

韻律制御手法としては、自然音声の観測を基に、人間が規則を構築するルールベース手法が考えられ、特に、現象を捉えやすい基本周波数（F_0）パターンについては、朗読調音声について良好な結果が得られている。ルールベース手法には、少量の分析結果に基づいて規則作成が可能と言う利点があるが、そのためにはエキスパートが必要であり、種々の調子の音声に規則を拡大することは必ずしも容易でない。このような観点から、韻律についてもコーパスベース手法によって生成することが行われている。このような手法の1つとして注目されるのは、HMMに基づく音声合成手法である。この手法では、音声の音韻的な特徴と韻律的な特徴を統合的に取り扱うため、両者の同期が自動的に取れるなどの利点がある。また、各フレームのF_0値をそのまま学習データとして用いるため、学習コーパスの作成は、容易という特長がある。合成班の小林、徳田らによって、感情音声や種々の調子の音声の合成が行われ、良好な結果が得られている。しかしながら、韻律に現われる発話構造を明示的に捉えるものではないため、我々の韻律に対する知識を反映させることは困難である。

これに対し、我々は、F_0パターンの生成過程のモデル（F_0モデル[1]）を用いたコーパスベース韻律生成手法の開発を進めた[2-11]。各フレームのF_0値でなく、F_0モデルの指令の大きさと時点を統計的手法の推定対象とすることにより、得られるF_0パターンには、モデルの制約が加わることになる。勿論、モデルの制約により、マイクロプロソディーなど、モデルで想定されていないF_0の動きは、基本的には表現することが出来ないが（モデルとの誤差を学習対象とすることにより表現可能）、反面、不自然なF_0の動きを排除することが可能となる。F_0モデルの指令は言語情報に直接対応しており、指令のレベルを変更して、感情のレベルを変化させると言ったことも可能となる。また、発話構造の変化を含め、発話スタイルによる韻律の変化を指令のレベルで明示的に捉えることが可能となり、例えば、俳優等のプロの感情表現に際する韻律制御を任意の話者に適用すると言ったことが実現できる。

開発した手法では、学習コーパスについてF_0モデルの指令を精度良く求めることが重要である。F_0モデルのパラメータ推定は直接解析的に求めることが出来ず、観測されたF_0パターンについて、モデルによる逐次近似（Analysis-by-Synthesis）を行って得るため、その初期値の設定が重要となる。これを人手で行なうことにより精度の良いパラメータ推定が可能となるが、これを種々の発話スタイルの多量の音声データについて行なうことは非現実的である。これに対しては、既に研究室で自動的にパラメータを抽出する手法 [12] を開発しているが、ある程度の抽出誤りは避けられず、それを用いて学習した統計的手法により推定されるF_0モデルパラメータに大きな誤りが含まれる結果となっていた。そこで、学習コーパスのテキストから得られる言語情報を利用し、パラメータの抽出精度を向上させることを行った。以下、言語情報を利用しない場合を旧(Original)手法[6-8]、利用した場合を新(New)手法[9-11]として区別する。

本手法は、種々の調子の音声に利用可能なものであるが、本研究では、感情音声を中心に研究を進めた。

2. F_0モデル

F_0モデルは、対数F_0の時間変化パターンが、句頭から句末に向う比較的緩やかな下降のフレーズ成分、語のアクセント型に対応した局所的な上昇下降のアクセント成分、さらに発話のF_0レベルに対応した基底周波数の和として表現されるとした上で、フレーズ成分とアクセント成分を、それぞれ固有の臨界制動2次線形系に対するインパルス応答、ステップ応答で近似するものである。インパルス関数はフレーズ指令、ステップ関数はアクセント指令と呼ばれ、例えば、前者は発話の構造、後者はアクセント型と対応するなど、言語情報、パラ言語・非言語情報との関連が、他のモデル化

よりもより直接的という、優れた特長を有する。

開発した手法での推定対象は、指令の大きさと時点である。時点については、対応するモーラ境界を基準点とする。F_0モデルのパラメータは、この他、臨界制動2次線形系の時定数等があるが、それらはほぼ一定値に固定し得ることが分かっており、推定対象とはしていない。

3. 韻律コーパス

本研究に用いた音声データは、1名の女性ナレータが予め用意された文リストを、怒り、喜び、悲しみの3感情と平静で読み上げたものである。平静はATR連続音声コーパス503文であるが、個々の感情については、読み上げる際にその感情を込めやすい文としてある。この音声データは奈良先端科学技術大学院大学の鹿野研究室で収録されたものである。音声データのうち、想定した感情が込められていないものを、簡単な聴取実験により排除した後、以下のプロセスにより、自動的に韻律コーパスを作成した（図1）。

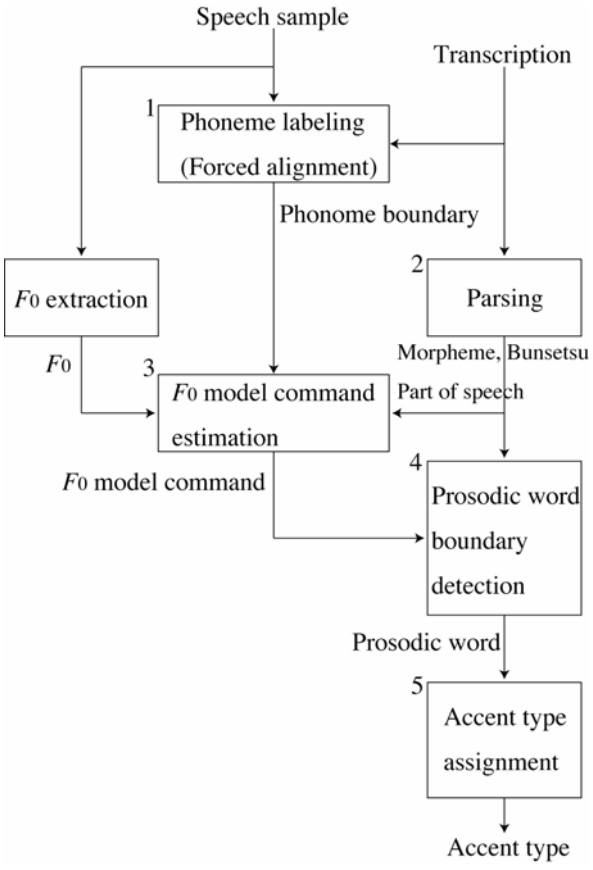

Fig. 1. Automatic production of prosodic corpus.

1. 音声認識エンジンJulius [13] を用いた音声波形への音素ラベリング。

2. 日本語形態素解析ソフト茶筅 [14] を用いた読み上げテキストの形態素の同定（と品詞情報の推定）。JUMAN とKNP [15] を用いたテキストの文節への分割と境界の深さの推定。文節境界の深さは、現在の文節が直接どの文節にかかるかのKNPコードとして与えられる。今回は、これらの言語情報を修正無しに次節のF_0モデルパラメータの推定に利用した。

3. 音声波形から抽出したF_0パターンから、研究室で開発した手法[12] によりF_0モデルパラメータを抽出。なお、旧手法では、抽出の際に特に発話テキストの言語情報から得られるF_0モデルパラメータに関する制約を利用していなかったが、そのため、抽出したパラメータに多くの誤差が含まれ、F_0モデルパラメータ推定の精度を下げる結果となっていた。そこで、新手法では、フレーズ指令は文節境界の前に位置する（境界が休止を伴う場合は100ms～300ms前、伴わない場合は0ms～100ms前）、2つのアクセント指令が近接し（50ms以下）なおかつ大きさに違いがない（0.1以下）場合は1つの指令に統合する、という制約の元でF_0モデルパラメータを抽出した。

4. 上記のプロセスで得られたアクセント指令をもとに、個々の文節境界が韻律語境界であるか否かを判定。1つの文節境界が2つのアクセント指令の間に位置する場合は、それを韻律語境界とし、複数の文節境界が位置する場合は、後の文節に近いものを韻律語境界とする。文節境界がない場合は、アクセント指令間で最も時間的に遅い形態素境界を韻律語境界とする。ここで、韻律語は、1つのアクセント成分を有する発話の単位でアクセント句とも呼ばれる。文節が長い複合語を有する場合は、2つあるいはそれ以上のアクセント成分を有するが、今回の音声データにはそのようなものは含まれていない。

5. このようにして得た個々の韻律語について、アクセント辞書を用いてアクセント型を付与。アクセント辞書は単語のアクセント型とともにアクセント結合の際の属性情報を有する。アクセント辞書を用いて実際にアクセント型を与える手法については研究室で開発したもの [16] を用いた。

新手法で付加したフレーズ指令に対する制約は、長い

複合語を含む場合には、誤りの原因となる可能性がある。これは、3語以上からなる複合語は、語内構造を持ち、それに対応して、文節内にフレーズ指令が生起することがあるためである。ただし、今回扱った音声データにはそのようなケースはない。

以上のプロセスにより、それぞれの感情について、400文程度の韻律コーパスを得た。これを表1に示すように学習データとテストデータに分け、F_0パターンの生成実験を行った。なお、3番目のF_0モデルパラメータ抽出のプロセスにおいて、基底周波数を個々の感情毎に、F_0の平均値から標準偏差の3倍を引いたものに固定した。具体的な値は、平静147.7 Hz, 怒り182.5 Hz, 喜び210.3 Hz, 悲しみ182.5 Hz である。この値は、推定されたF_0モデルパラメータからF_0パターンを作成する際にも用いた。

Table 1. Number of samples used for the experiment.

Type	Category	Number	
		Sentence	Prosodic word
Calm	Training	333	2340
	Testing	50	338
Anger	Training	472	3247
	Testing	50	346
Joy	Training	358	2391
	Testing	50	271
Sadness	Training	305	2185
	Testing	50	389

4. F_0パターン生成

旧手法ではF_0モデルパラメータの推定は、韻律語毎に行ない、総ての韻律語についてパラメータの推定が終了した後に、文全体のF_0パターンを生成していた。このため、テキストを入力とした場合、まず韻律語境界を決定する必要がある。フレーズ指令位置に関する制約を課していないため、アクセント指令中にフレーズ指令が生起するなどのF_0モデルで想定していない推定結果が得られるなどの問題があった。

これに対し、新手法では、まず、フレーズ指令を推定する。その際、フレーズ指令は文節境界のみで生起するという制約を課す。この制約は、3節の韻律コーパス作成に際してフレーズ指令位置に課した制約と対応する。これにより、旧手法での問題が解消される。テキストを入力として、次の4つのプロセスにより、F_0モデルのパラメータを推定する。

1. フレーズ指令の 推定。個々の文節境界について、まず、そこにフレーズ指令が生起するか否かを推定する。次に、生起する場合は、指令の大きさと時点を推定する。（今回の実験では、次のプロセスで韻律語境界でないと判定された文節境界は推定対象としていない。）

2. 韻律語境界の推定。各形態素境界について、韻律語境界であるか否かを推定する。

3. アクセント型の決定。各韻律語のアクセント型を3節の5のプロセスと同様にして決定する。

4. アクセント指令の推定。各韻律語に対し、アクセント指令の大きさと時点を推定する。

上記において、1, 2, 4 のプロセスは２分決定木を用いて行った（Edinburgh Speech Tools Library の CART [http://www.cstr.ed.ac.uk/projects/speech_tools/] を利用）。統計的手法としては、他に、多重線形回帰式やニューラスネットワークによるものが考えられるが、性能に大きな差がない。以下に、各プロセスを説明する。なお、3 は日本語のアクセント型の規則から直接に韻律語のアクセント型を求めるプロセスで、統計的な推定ではない。

4.1 フレーズ指令の推定

フレーズ成分推定の２分決定木への入力パラメータを表2に示す。問題としている文節境界に先行・後続する文節の言語情報の他、境界の統語的な深さを表わす Boundary Depth Code (BDC)を入力パラメータとした。表の括弧中のカテゴリ数は先行文節に対してであり、先行文節がない（文頭）ということを示すために、後続文節に対するカテゴリ数より1だけ大きい。BDC は、KNP コードから簡単に求めることが出来る。図2は「あらゆる現実を総て自分のほうにねじ曲げたのだ」の例の例であるが、掛かり先の文節の番号から現在の文節の番号を引いた"Distance"を右に1つシフトすると得られる。

２分決定木の出力パラメータは、フレーズ指令の有無を表わすフラグ PF と「有り」の場合の指令の大きさと時点である。PF は先行研究の結果に基づいて導入した。

Table 2. Input parameters for phrase command prediction. The category numbers in the parentheses are those for the directly preceding bunsetsu.

Input parameter	Category
Position in sentence	28
Number of *morae*	21 (22)
Accent type (location of accent nucleus)	18 (19)
Number of words	10 (11)
Part-of-speech of the first word	14 (15)
Conjugation form of the first word	19 (20)
Part-of-speech of the last word	14 (15)
Conjugation form of the last word	16 (17)
Boundary depth code (BDC)	20
Phrase command for preceding *bunsetsu*	2
Number of morae between the preceding phrase command and the head of the current *bunsetsu*	25
Magnitude of the preceding phrase command	Continuous

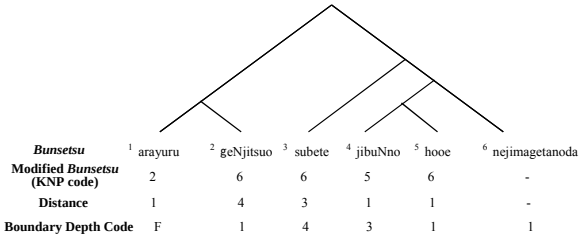

Fig. 2. Result of syntactic analysis by KNP and bunsetsu boundary depth codes for the sentence "arayuru geNjitsuo subete jibuNno hooe nejimagetanoda ([He] twisted all the reality to his side.)." Code F denotes the prosodic word locating at the sentence initial. The code takes the value 1 when the boundary between the current and preceding bunsetsu's is left branching.

Table 3. Result of phrase command flag PF prediction. (in %)

	Closed	Open
Calm	67.4	64.9
Anger	69.0	66.1
Joy	66.1	63.2
Sadness	74.5	74.8

表3 は、PF 推定結果の正解率である。「悲しみ」で若干高い値が得られている。なお、音声データの休止の情報を用いると、正解率が若干向上するが、これは休止をフレーズ指令に先立って推定して利用することの可能性を示唆している。また、表4 は、推定されたフレーズ指令の大きさのターゲット値との RMS 誤差である。

Table 4. Root mean square errors of phrase command magnitude prediction.

	Closed	Open
Calm	0.229	0.228
Anger	0.162	0.192
Joy	0.168	0.155
Sadness	0.144	0.131

実際に観測される基本周波数パターンでは、大きなフレーズ指令が隣接して先行する場合、当該境界のフレーズ指令は小さいかあるいは存在しないことが多い。この現象に対応して、表2 の下3段のパラメータを入力パラメータに加えている。しかしながら、今回の実験では、その効果ははっきりしない。この1因として、先行指令に含まれる推定誤りが考えられる。

4.2 韻律語境界の推定

表5 に示すように、問題としている形態素境界の先行・後続形態素の言語情報と4.1で推定されたフレーズ指令の情報を入力パラメータとした。推定される出力パラメータは、形態素境界が韻律語境界か否かの2値のフラグである。表6 に示すように、総ての場合について、80%を超える精度で推定が行われている。

Table 5. Input parameters for prosodic word boundary prediction. The category numbers in the parentheses are those for the directly preceding morpheme.

Input parameter	Category
Part-of-speech	15 (16)
Conjugation form	24 (25)
Conjugation type	35 (36)
Number of *morae*	9 (10)
Position in sentence	63
BDC of *bunsetsu* where the current morpheme belongs	22
Bunsetsu boundary at the head of the current morpheme	2
PF for current morpheme	2
Number of morae between the preceding phrase command and the head of the current morpheme	31
Magnitude of the preceding phrase command	Continuous

Table 6. Result of prosodic word boundary prediction. (in %)

	Closed	Open
Calm	88.5	87.7
Anger	87.0	83.7
Joy	86.7	85.3
Sadness	85.3	85.0

4.3 アクセント指令の推定

表7 のようにフレーズ指令推定の場合と同様の入力パラメータでアクセント指令推定を行った。出力パラメータはアクセント指令の大きさと時点である。表8に大きさ推定の誤差を RSM 値で示す。「悲しみ」で小さな値となっているが、これは、指令の大きさ自体が他の場合より小さいからである。以前の文音声の基本周波数パターンの分析結果から、フレーズ指令とアクセント指令の大きさには負の相関があること、先行するアクセント指令の位置と大きさが、後続するアクセント指令の大きさに関係があることが知られている。これに対応して、当該フレーズ指令と先行アクセント指令のパラメータを入力パラメータに加えることも行ったが、ほとんど結果には影響なかった。これも、推定結果に含まれる誤りが原因と考えられる。

Table 7. Input parameters for accent command prediction. The category numbers in the parentheses are those for the directly preceding prosodic word.

Input parameter	Category
Position in sentence	27
Number of *morae*	17 (18)
Accent type (location of accent nucleus)	16 (17)
Number of words	9 (10)
Part-of-speech of the first word	14 (15)
Conjugation form of the first word	23 (24)
Part-of-speech of the last word	14 (15)
Conjugation form of the last word	23 (24)
Boundary depth code	22

Table 8. Root mean square errors of accent command amplitude prediction.

	Closed	Open
Calm	0.158	0.170
Anger	0.162	0.181
Joy	0.153	0.130
Sadness	0.112	0.127

4.4 F_0パターン生成

図3 は、平静の「親父は頑固だけれどもそんなえこひいきはせぬ男だ」について、それぞれ新手法と旧手法で予測された指令で生成したF_0パターンをターゲットのF_0パターンと比較して示したものである。ここでターゲットのF_0パターンは、観測されたF_0パターンから抽出した指令で生成したパターンである。明らかに、新手法で生成したF_0パターンの方がターゲットに近い。特に、旧手法で見られたフレーズ指令の位置に対する誤りが解消されている。

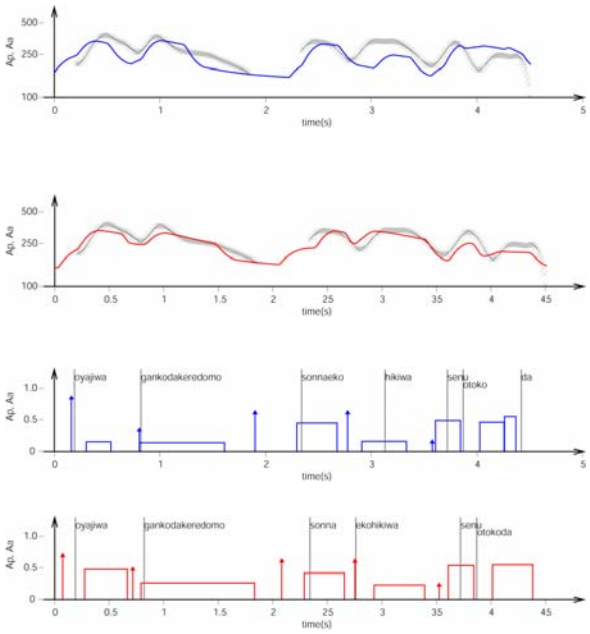

Figure 3. F_0 contours generated by the original method (first panel) and the new method (second panel) for calm speech "oyajiwa gaNkodakeredomo soNna ekohiikiwa senu otokoda (Although my dad is a tough person, he never shows such a prejudice to any person.)." The third and the fourth panes show the model commands extracted for the target F_0 contour and those predicted by the new method, respectively. The thick and half tone contour in the first and the second panel is the target F_0 contour.

指令の推定結果を総合的に評価する指標として、推定された指令で生成されたF_0パターンとターゲットのF_0パターンの平均二乗誤差F_0MSEを次式で定義して用いる。

$$F_0MSE = \frac{\sum_t (\Delta \ln F_0(t))^2}{T}, \qquad (1)$$

ただし、$\Delta \ln F_0(t)$ は、時刻tにおける、対数尺での、2つのF_0パターン間の差である。和はF_0が観測される有

声区間のみについてとる。Tは文におけるこのような有声区間の総フレーム数である。得られたF_0*MSE*の平均を、学習に用いた音声データについて、また学習に用いない評価用データについて求めた結果を表9 に示す。「悲しみ」で最も誤差が小さいという結果が得られているが、先に述べたように指令の大きさ自体が小さいためと考えられる。

Table 9. Average F_0MSE's of F_0 contours generated using the predicted model parameters.

	Closed	Open
Calm	0.049	0.048
Anger	0.051	0.065
Joy	0.052	0.078
Sadness	0.035	0.043

5. 感情音声合成と聴取実験

推定したF_0モデルの指令で生成したF_0パターンを用いて音声合成を行った。合成に際して、各音素の長さを決める必要があるが、これについては、指令の推定と同様な手法で行った[6]。分節的特徴は情報処理振興事業協会(IPA)独創的情報技術育成事業の「擬人化音声対話エージェント基本ソフトウエアの開発（代表：嵯峨山茂樹)」で用意されたHMM音声合成ツールキットを用いた。Tri-phoneモデルの学習には表1に示した学習データを用いた。特徴量はケプストラム係数24次元とそのΔ及びΔ^2値からなる75次元である。サンプリング周波数、フレーム間隔、フレーム長は、それぞれ16 kHz、5 ms、25 msである。感情毎に、表1のテストデータから10文を適宜選択して音声合成を行った。9名の日本語話者に対して合成音声を提示し、平静、怒り、喜び、悲しみの内から強制選択させた。表10に結果を示す。さらに、同じ被験者に対して、どの程度、感情を感じ取れたか(5: 大変良く, 3: 普通, 1: ほとんど感じない)、合成音声の自然性はどうか(5: 自然, 3: まあまあ, 1: 非常に合成音的)、について尋ねた。表11に示すように、前者に関しては、「怒り」で良好な結果が得られた。「喜び」、「悲しみ」では若干評価が低下した。自然性に関しては、低い評価であったが、これは、HMM合成による分節的特徴に起因する品質の劣化が関係していると考えられる。

Table 10. Percentages showing how correctly the designated emotion (anger, joy, sadness) in synthetic speech is perceived. The italic numbers indicate the percentages when the designated emotion is perceived correctly. "Ori." indicates the results when the commands predicted by the original method are used, while "New" indicates those predicted by the new method are used. The results are averaged over all 10 sentences and 9 speakers for each emotion.

	Anger		Joy		Sadness	
	Ori.	New	Ori.	New	Ori.	New
Calm	10.0	7.8	30.0	23.3	32.2	26.7
Anger	78.3	83.3	11.1	10.0	11.7	10.0
Joy	6.1	5.6	56.7	57.8	11.7	7.8
Sadness	5.6	3.3	2.2	8.9	44.4	55.6

Table 11. Scores for the realization of the designated emotion and naturalness of prosody.

	Anger		Joy		Sadness	
	Ori.	New	Ori.	New	Ori.	New
Degree	4.01	4.21	3.26	3.36	3.07	3.12
Quality	2.06	2.48	1.76	1.90	1.61	2.32

6. 結論と今後の予定

テキストを入力として感情音声合成のためのF_0パターンを生成するコーパスベース手法を開発した。この手法は、F_0の値を直接推定する代わりにF_0モデルの指令を推定するもので、F_0モデルの制約により、推定が良好に行かない場合でもある程度の品質が得られると言う特徴がある。学習にはF_0モデルの指令が表記されている韻律コーパスを用いるが、この指令を、言語情報を利用して高精度で自動抽出する。HMM音声合成により得られた合成音声の聴取の結果、言語情報を利用して韻律コーパスを作成することの有効性が確認された。

F_0モデルの最大の利点は、生成されたF_0パターンと言語情報等の各要因との対応が明確であることである。同一の文について、感情音声と平静音声のF_0パターンを求め、指令レベルでの比較を行なうことにより、平静音声からどのようにすれば感情音声を得ることができるかの知見を得ることが可能である。特別な訓練のない話者にとって、種々のスタイルの音声を適切に発声することは困難と考えられるが、それを可能とする。

参考文献

[1] H. Fujisaki, H. and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," Journal of Acoust. Soc. Japan, Vol.5, No.4, pp.233-242 (1984-10).

[2] A. Sakurai, K. Hirose, and N. Minematsu, "Data-driven generation of F_0 contours using a superpositional model," Speech Communication, Vol.40, No.4, pp.535-549 (2003-6).

[3] K. Hirose, M. Eto, N. Minematsu, and A. Sakurai, "Corpus-based synthesis of fundamental frequency contours based on a generation process model," Proc. European Conference on Speech Communication and Technology, Aalborg, pp.2255-2258 (2001-9).

[4] K. Hirose, N. Minematsu, and M. Eto, "Data-driven synthesis of fundamental frequency contours for TTS systems based on a generation process model," Proc. International Conference on Speech Prosody, Aix-en-Provence, pp.391-394 (2002-4).

[5] K. Hirose, M. Eto, and N. Minematsu, "Improved corpus-based synthesis of fundamental frequency contours using generation process model," Proc. International Conf. on Spoken Language Processing, Denver, pp.2085-2088 (2002-9).

[6] K. Hirose, T. Katsura, and N. Minematsu, "Corpus-based synthesis of F_0 contours for emotional speech using the generation process model," Proc. International Congress of Phonetic Sciences, Barcelona, pp.2945-2948 (2003-8).

[7] K. Hirose, T. Ono, and N. Minematsu, "Corpus-based synthesis of fundamental frequency contours of Japanese using automatically-generated prosodic corpus and generation process model," Proc. European Conference on Speech Communication and Technology, Geneva, pp.333-336 (2003-9).

[8] K. Hirose, K. Sato, and N. Minematsu, "Emotional speech synthesis with corpus-based generation of F_0 contours using generation process model," Proc. International Conference on Speech Prosody, Nara, pp.417-420 (2004-3).

[9] K. Hirose, K. Sato, and N. Minematsu, "Corpus-based generation of F_0 contours using generation process model for emotional speech synthesis," Proc. International Workshop, From Sound to Sense: 50+ Years of Discoveries in Speech Communication (CD-ROM), Boston, pp.C37-C42 (2004-6)

[10] K. Hirose, K. Sato, and N. Minematsu, "Corpus-based synthesis of fundamental frequency contours with various speaking styles from text using F_0 contour generation process model" Proc. ISCA Speech Synthesis Workshop (CD-ROM), Pittsburgh, pp.162-166 (2004-6).

[11] K. Hirose, K. Sato, and N. Minematsu, "Improvement in corpus-based generation of F_0 contours using generation process model for emotional speech synthesis," Proc. International Conference on Spoken Language Processing, Jeju, pp.1349-1352 (2004-10).

[12] 成澤修一, 峯松信明, 広瀬啓吉, 藤崎博也, "音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, pp.2155-2168 (2002-7).

[13] A. Lee, T. Kawahara, and K. Shikano, "Julius – an open source real-time large vocabulary recognition engine," Proc. European Conf. on Speech Communication and Technology, Aalborg, pp.1691-1694 (2001).

[14] 松本裕治, "形態素解析システム「茶筌」," 情報処理, Vol.41, No.11, pp.1208-1214 (2000).

[15] S. Kurohashi and M. Nagao, "A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures," Journal of Computational Linguistics, Vol.20, No.4, pp.507-534. (1994).

[16] N. Minematsu, R. Kita, K. Hirose, "Automatic estimation of accentual attribute values of words for accent sandhi rules of Japanese text-to-speech conversion," IEICE Trans. Information and Systems Vol. E86-D, No.3, pp.550-557 (2003-3).