音声 F_0 パターン生成過程の確率モデルに基づく非母国語話者音声の韻律補正法の検討*

☆門脇健人 [1], 大石康智 [2], 石原達馬 [1], 北条伸克 [1], 亀岡弘和 [1,2]
([1] 東大院・情報理工, [2]NTT CS 研)

1 はじめに

ネットワークを利用した高速・高品質な音声通信手段が普及する中，母国語以外の言語で他国の人々とコミュニケーションを図る機会が増えている。しかしながら，たとえ文法的に正しい内容を発話したとしても，対象とする言語の発音や抑揚を再現できず，内容を相手に正しく理解してもらえないという経験がだれにでもあるのではないだろうか。特に日本語を母国語とする話者は，英語学習教育を幼少期から受けているものの，英語の発声を正しく修得して活用することが難しい。この原因の一つは，言語の韻律的特徴が互いに異なるためと言える。韻律とは，声の強弱や長短，高低，リズムを指し，例えば，日本語はモーラ拍リズム型，英語は強勢拍リズム型，フランス語やヒンズー語は音節拍リズム型で韻律が変化する。中国語やタイ語，ベトナム語では，1 音節内で声の高低が変わる音節声調をもつことも特徴と言える [1]。このような言語特有の韻律的特徴のため，母国語以外の新しい言語を修得することが難しいと考えられる。

本研究では，様々な言語の韻律的特徴を考慮し，言語教育支援やコミュニケーションの円滑化を目的として，非母国語話者の韻律を母国語話者らしい韻律へと自動的に補正する手法を提案する。具体的には，音声基本周波数の時間変化（F_0 パターン）を韻律的特徴とみなし，F_0 パターンの生成過程を表現する藤崎モデル [2] を韻律補正に応用する。藤崎モデルは，F_0 パターンをフレーズ成分とアクセント成分の和で表す物理モデルであり，それぞれ甲状軟骨の並進運動による緩やかな時間変化と回転運動による急激な時間変化に対応すると仮定する。このフレーズ成分とアクセント成分の使い方が言語によって異なると想定する。我々はこれまで，藤崎モデルを基礎とした F_0 パターンの生成過程の確率モデルを提案し，観測される F_0 パターンからフレーズ成分とアクセント成分を推定する逆問題を検討した [3, 4]。さらに，有限種類のアクセント型を反映した語彙テンプレートの連結によって F_0 パターンが生成されるという統計的語彙モデル [5] を提案し，語彙の学習を試みた。背後に存在する有限種類の語彙テンプレートから F_0 パターンが生成されるという仮説の妥当性を示すことができた。本稿ではこれらの結果に基づいて，英語に特有な有限種類の語彙テンプレートを学習し，日本語母語話者による英語発声を，英語の語彙テンプレートに置き換えて F_0 パターンを補正する。Fig. 1 に示すように，音声分析合成法 TANDEM-STRAIGHT[6] を利用して，入力音声をあらかじめ F_0 パターンとスペクトル包絡，非周期成分に分解し，F_0 パターンだけを補正して再合成する。聴取実験を通して，再合成された音声を，韻律の聞き取りやすさの観点から評価し，提案法の有効性を検証する。

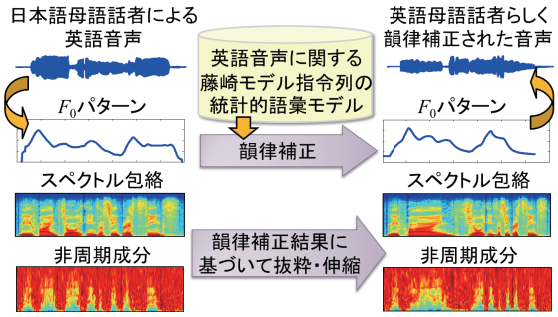

Fig. 1 非母国語話者音声の韻律補正

非負値時空間分解法 [7] を用いて日本語母語話者が発声した英語の発話リズムを英語母語話者らしく変換する手法が検討されているが，ここでは同一文章を発声したパラレルデータを必要とした。言語情報を用いずに，大量の F_0 パターンから，言語のフレーズとアクセントのテンプレートを学習して，音声の韻律を補正することが本研究の特徴的な点である。

2 藤崎モデル指令列の統計的語彙モデル

藤崎モデル [2] では，対数 F_0 パターン $y(t)$ が以下のように 3 つの成分の和で表される。

$$y(t) = x_{\mathrm{p}}(t) + x_{\mathrm{a}}(t) + x_{\mathrm{b}} \qquad (1)$$

ここで，t は時間，$x_{\mathrm{p}}(t)$ はフレーズ成分，$x_{\mathrm{a}}(t)$ はアクセント成分，x_{b} はベースライン成分（時間によらない定数）である。さらにフレーズ成分とアクセント成分はそれぞれ，デルタ列からなるフレーズ指令 $u_{\mathrm{p}}(t)$ と矩形パルス列からなるアクセント指令 $u_{\mathrm{a}}(t)$ の 2 次系フィルタの出力で表現される。それぞれの

* Automatic prosodic correction method for non-native speaker based on probabilistic model of speech F_0 contour. by KADOWAKI, Kento (The University of Tokyo), OHISHI, Yasunori (NTT), ISHIHARA, Tatsuma, HOJO, Nobukatsu, KAMEOKA, Hirokazu (The University of Tokyo)

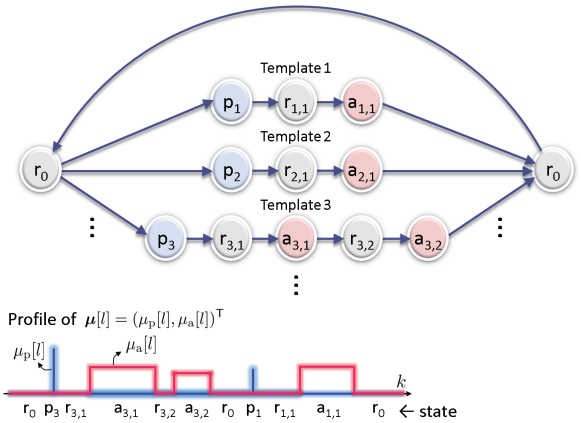

Fig. 2　統計的語彙モデルに基づくフレーズ・アクセント指令列の状態遷移トポロジー

2 次系フィルタの応答の速さを表す角周波数パラメータを α，β とし，個人や発話によらずおおよそ $\alpha = 3$ rad/s，$\beta = 20$ rad/s 程度の値となることが知られている。

2.1　藤崎モデルの確率モデル化

藤崎モデルにおけるフレーズ指令とアクセント指令はそれぞれ，デルタ列と矩形パルス列で表現され，さらにこれらは互いに重ならないという仮定が置かれる。文献 [3, 4] では，これらの制約を満たす指令列を確率モデルとして記述するために，指令のペア $\boldsymbol{o}[k] = (u_p[k], u_a[k])^T$ を，隠れマルコフモデル (HMM) の出力として表現した。各状態の出力分布を正規分布と仮定すると，出力は

$$\boldsymbol{o}[k] \sim \mathcal{N}\left(\boldsymbol{o}[k]; \boldsymbol{c}_{s_k}, \boldsymbol{\Upsilon}_{s_k}\right) \tag{2}$$

に従う。ここで s_k は時刻 k における状態を表す。上式の平均 $\boldsymbol{c}_{s_k}$ と分散 $\boldsymbol{\Upsilon}_{s_k}$ は HMM の状態遷移によって得られる出力分布のパラメータからなる。実際に観測される F_0 パターン $\boldsymbol{y} = \{y[k]\}_{k=1}^K$ は $\{\boldsymbol{o}[k]\}_{k=1}^K$ の各指令に 2 次系フィルタが畳み込まれ，さらにベースライン成分 u_b を足したものを平均とし，$v_n^2[k]$ を分散パラメータとする確率密度関数

$$P(\boldsymbol{y}|\boldsymbol{o}) = \prod_{k=1}^{K} \mathcal{N}(y[k]; x[k], v_n^2[k])$$
$$x[k] = G_p[k] * u_p[k] + G_a[k] * u_a[k] + u_b \tag{3}$$

から生成される。ここで，G_p と G_a はフレーズ成分とアクセント成分を生成する 2 次系フィルタを表す。

2.2　指令列の統計的語彙モデル

HMM を利用した藤崎モデルの確率モデル化では，言語学的な先験知識を考慮していなかったが，本来はイントネーション型などに起因する言語的な制約によって指令列のとりうる範囲を制限するべきである。通常の発話において，イントネーション型が限られる

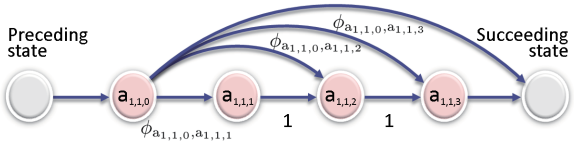

Fig. 3　継続時間長を表現するための状態分割: 遷移確率 $\phi_{a_{1,1,0},a_{1,1,1}}$ は状態 $a_{1,1}$ が 4 回継続する確率に対応する。

ことや，違う発話内容であっても同一のイントネーションを持つものが存在する。これは，藤崎モデルの指令列が有限種類の語彙テンプレートの連結によって表現できる可能性を示唆する。文献 [5] では，そのような有限の語彙テンプレートを，Fig. 2 のようなテンプレート間を遷移可能な Left-to-Right 型 HMM で表現した。ここで，状態の継続時間長も重要な情報であるため，Fig. 3 のような状態分割により継続時間をモデル化する。これによってテンプレート内の休息やアクセント指令の継続時間を表現することができる。推定すべきパラメータはフレーズ指令およびアクセント指令の大きさ（HMM の各状態の平均パラメータ），指令列を生成する分散パラメータ，実際の指令列 $\{\boldsymbol{o}[k]\}_{k=1}^K$，そして HMM の状態遷移確率である。補助関数法を用いて，パラメータの事後確率を最大化するようにパラメータは反復して推定される。

3　藤崎モデル指令列の統計的語彙モデルを利用した韻律補正

藤崎モデル指令列の統計的語彙モデルを利用して，日本語母語話者による英語音声の韻律を英語母語話者らしい韻律へと補正して音声を再合成する手法を提案する。この手法は，英語音声の F_0 パターンを生成するために必要な語彙テンプレートを学習するステップと，入力となる日本語母語話者による英語発声の F_0 パターンを補正するステップからなる。

3.1　英語音声の統計的語彙モデルの学習

藤崎モデル指令列の統計的語彙モデルによって，日本語の F_0 パターンが有限個の語彙テンプレートの連結によって表現されることが示された [5]。日本語音声と英語音声における韻律のモーラ拍リズム型と強勢拍リズム型の違いが，この統計的語彙モデルで表現されることを仮定し，複数の英語音声の F_0 パターンを用いて，前節の語彙テンプレートを学習する。

3.2　入力音声の韻律補正と音声の再合成

日本語母語話者による英語発声の F_0 パターンを補正する手順を説明する。

1. 音声分析合成手法 TANDEM-STRAIGHT[6] を利用して，入力音声信号から F_0 パターン，スペクトル包絡，非周期成分を抽出する。

2. 抽出された F_0 パターン $\{\boldsymbol{y}[k]\}_{k=1}^{K}$ を，3.1 節で学習された統計的語彙モデルの観測データとみなし，Forward-Backward アルゴリズムによって HMM の状態系列が周辺化されて得られる指令列を $\{\boldsymbol{o}_{\mathrm{input}}[k]\}_{k=1}^{K}$ とする。

3. 同様に，$\{\boldsymbol{y}[k]\}_{k=1}^{K}$ を，3.1 節で学習された統計的語彙モデルの観測データとみなし，Viterbi アルゴリズムによって得られる最尤な状態系列を $\{s_{\mathrm{input}}[k]\}_{k=1}^{K}$ とする。

4. $\{s_{\mathrm{input}}[k]\}_{k=1}^{K}$ の中で，自己遷移によって，ある状態に停留する区間に対し，自己遷移確率を使って計算される状態継続長分布の平均値がその停留時間（継続長）となるように伸縮を行う。伸縮された新たな状態系列を $\{s_{\mathrm{output}}[k]\}_{k=1}^{K'}$ とする。

5. $\{s_{\mathrm{output}}[k]\}_{k=1}^{K'}$ と HMM の平均パラメータを使って，指令列 $\{\boldsymbol{o}_{\mathrm{output}}[k]\}_{k=1}^{K'}$ を作成する。

6. $\{\boldsymbol{o}_{\mathrm{input}}[k]\}_{k=1}^{K}$ と $\{\boldsymbol{o}_{\mathrm{output}}[k]\}_{k=1}^{K'}$ それぞれに，式 (3) のフレーズ成分およびアクセント成分を生成するための 2 次系フィルタを畳み込み，再合成された F_0 パターン $\{x_{\mathrm{input}}[k]\}_{k=1}^{K}$ と韻律補正された F_0 パターン $\{x_{\mathrm{output}}[k]\}_{k=1}^{K'}$ を計算する。

7. 韻律補正された音声を再合成するために，動的計画法（DP マッチング）を利用して，$\{x_{\mathrm{input}}[k]\}_{k=1}^{K}$ と $\{x_{\mathrm{output}}[k]\}_{k=1}^{K'}$ の最適マッチング経路を計算する。

8. 最適経路に従って入力音声のスペクトル包絡と非周期成分を抜粋し，韻律補正された音声を TANDEM-STRAIGHT で再合成する。

DP マッチングでは，入力信号 $I_i(1 \le i \le I)$ と参照信号 $R_j(1 \le j \le J)$ の最適なマッチング経路を求める。以下のように局所的な傾斜を 1/2 と 2 の間に制限した漸化式を利用して，信号間の距離を求める。

$$
\begin{aligned}
&g(I_i, R_j) \\
&= \min \begin{cases} g(I_{i-2}, R_{j-1}) + 2d(I_{i-1}, R_j) + d(I_i, R_j) \\ g(I_{i-1}, R_{j-1}) + 2d(I_i, R_j) \\ g(I_{i-1}, R_{j-2}) + 2d(I_i, R_{j-1}) + d(I_i, R_j) \end{cases}
\end{aligned}
$$

ここで，$g(I_1, R_1) = 2d(I_1, R_1)$，$j = 1$ として，i を変えながら上式を計算し，次に j を増加させて $j = J$ となるまで同様の計算を繰り返す。局所距離は $d(I_i, R_j) = |I_i - R_j|$ とする。上式の最小値選択がいずれであったかを蓄えておき，(I, J) からこれを逆にたどることによって最適マッチング経路を求める。$\{x_{\mathrm{output}}[k]\}_{k=1}^{K'}$ と $\{x_{\mathrm{input}}[k]\}_{k=1}^{K}$ をそれぞれ，入力信号と参照信号として計算した例を Fig. 4 に示す。

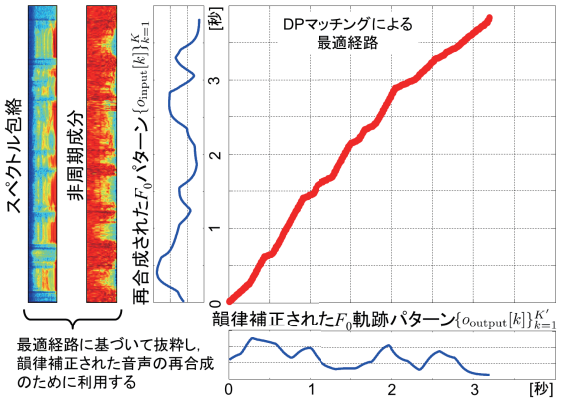

Fig. 4　再合成された F_0 パターンと韻律補正された F_0 パターンの DP マッチング

4　評価実験

提案法に基づく日本語母語話者による英語音声の韻律補正法の有効性を評価する。英語を母国語とする話者 4 名がそれぞれ，12 文章を発声した音声信号（合計 48 文章）を学習データとする。一方，英語を母国語としない日本人話者 1 名が上記の文章のうちランダムに選択された 5 つ文章を英語で発声した音声信号を評価データとする。学習データを用いて，英語音声の藤崎モデル指令列の統計的語彙モデルを学習する。F_0 は 8 ms ごとに推定され，2 次系フィルタの時定数は $\alpha = 3.0$ rad/s と $\beta = 20.0$ rad/s，フレーズ指令の分散パラメータは $\upsilon_{\mathrm{p}}^2[k] = 3^2$，アクセント指令の分散パラメータは $\upsilon_{\mathrm{a}}^2[k] = 0.03^2$，ベースライン成分の分散パラメータは $\upsilon_{\mathrm{b}}^2 = 10^{-8}$，有声区間のノイズ成分の分散パラメータは $\upsilon_{\mathrm{n}}^2[k] = 10^{15}$，無声区間のノイズ成分の分散パラメータは $\upsilon_{\mathrm{n}}^2[k] = 0.1^2$ とし，これらはすべて固定した。ベースライン成分の平均 μ_{b} は F_0 パターンの有声区間の最低値に設定した。Fig. 2 における語彙テンプレート数は 8 つに固定し，アクセントの数が 1 つのテンプレートを 3 つ，アクセントの数が 2 つのテンプレートを 3 つ，アクセントの数が 3 個のテンプレートを 2 つとした。これらの状態遷移確率および，2 節で説明したモデルパラメータが補助関数法によって推定される。パラメータ推定の反復回数は 20 回とした。

Fig. 5 と Fig. 6 は日本語母語話者による英語音声の韻律が提案法によって補正された結果を示す。(a) が実際に発声した音声の F_0 パターンであり，(b) が 2 節の統計的語彙モデルによって再合成された F_0 パターン，(c) が 3 節に示す手順によって韻律補正された F_0 パターンを示す。統計的語彙モデルによる F_0 パターンの分析合成性能はまだ十分ではないが，おおよそ (a) と (b) の F_0 パターンは近いものとみなせる。さらに，(b) と (c) を比較した場合，単純に F_0 パターンを伸縮したものではないことがわかるだろう。

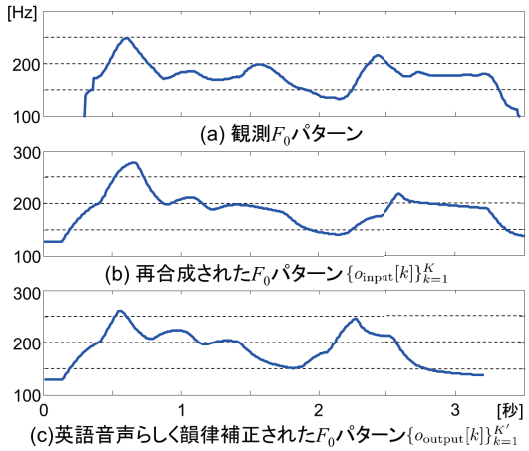

Fig. 5　日本語母語話者音声の韻律補正結果 1

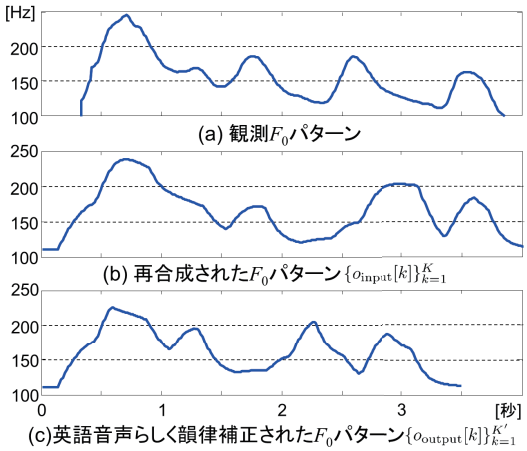

Fig. 6　日本語母語話者音声の韻律補正結果 2

英語母語話者による音声を使って学習された統計的語彙モデルにおけるフレーズやアクセントの大きさ，またアクセントや休止の継続時間長が反映されて，F_0 パターンが補正されることを確認できる。

次に韻律補正された音声の主観評価実験を行った。まず，原音声を被験者に聞いてもらい，その後，補正された音声を聞いて，補正された音声がより英語らしい韻律に変化したか否かを 5 段階で評価してもらった。3 はあまり変化がない，3 より高いと英語らしくなったという基準の下で，被験者 10 名がそれぞれ，5 つの文章を評価した。評価実験結果を Fig. 7 に示す。帰無仮説 $\mu = 3.0$，対立仮説 $\mu > 3.0$ として t-検定を行なったところ，文章 No.1，2，5 の 3 つにおいて有意水準 5%で帰無仮説が棄却された。すなわち，5 つの文章のうち，3 つは補正された音声が英語らしくなったと統計的に確認できた。ただし，実験が小規模であるため，学習データおよび評価データの量を増やし，大規模に提案法を評価することが今後の課題である。また，今回の実験では文章の内容に関してクローズドな評価であるため，学習データに含まれない文章を発声したときに，韻律補正がどの程度可能であるか，性能を評価することも必要である。

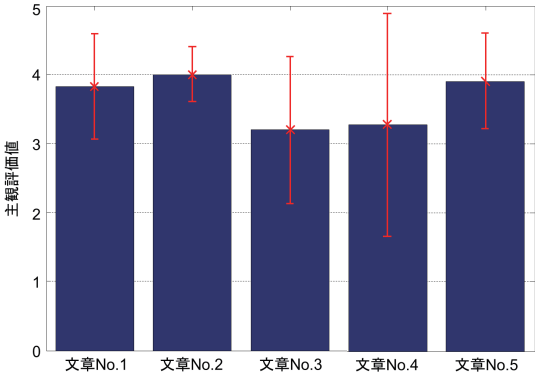

Fig. 7　5 段階評価による主観評価実験結果

5　おわりに

本稿では，言語教育支援やコミュニケーションの円滑化を目的として，日本語母語話者による英語音声を英語母語話者らしい韻律に補正して再合成する手法を提案した。音声の F_0 パターンを韻律的特徴とみなし，先行研究によって構築された藤崎モデル指令列の統計的語彙モデルを利用して，英語音声における典型的な指令列パターンを語彙テンプレートとして学習し，入力音声の F_0 の指令列パターンを英語の語彙テンプレートに置き換えて，F_0 パターンを補正する。提案法は F_0 パターンを単純に伸縮するのではなく，英語音声のフレーズやアクセントの大きさ，またアクセントや休止の継続長を反映して補正することが特徴である。主観評価実験より，一部の入力音声は英語らしく補正され，韻律の観点から聞き取りやすくなることを確認した。今後の課題は，学習データおよび評価データの量を増やし，大規模に提案法を評価することである。その中で，語彙テンプレートの数や HMM のトポロジーを調整することも必要となる。今回，評価に利用した日本語母語話者の英語の発声力は中級レベルである。英語の発声力に対して，提案法がどこまで韻律補正を可能とするか検証することも興味深い。また，提案法を英語以外の言語の韻律補正に応用することも現在検討中である。

参考文献

[1] H. Li *et. al.*, in *Proc. IEEE 2013.*
[2] H. Fujisaki, “In Vocal Physiology: Voice Production, Mechanisms and Functions,” Raven Press, 1988.
[3] H. Kameoka *et. al.*, in *Proc. SAPA 2010.*
[4] K. Yoshizato *et. al.*, in *Proc. Interspeech 2012.*
[5] T. Ishihara *et. al.*, in *Proc. Interspeech 2013.*
[6] H. Kawahara *et. al.*, in *Proc. ICASSP 2008.*
[7] S. Hiroya *et. al.*, IEEE Trans. ASLP, Vol.21, No.10, pp.2108–2117, 2013.