韻律指令列推定のための基本周波数・音韻特徴量系列同時生成モデルの検討 *

☆佐藤遼太郎 (東大), 亀岡弘和，柏野邦夫 (NTT)

1　はじめに

音声 F_0 軌跡の藤崎モデル指令列推定は，韻律の表現や変換等の応用上重要な問題である．近年，藤崎モデルに基づいた F_0 軌跡生成モデルを構築し，これに基づいて指令列の統計的推定を行う手法が研究されている．

本稿では，推定精度の向上を目的として，F_0 指令列推定に音韻特徴量の情報を手がかりとして用いる推定手法を検討し，実験により提案手法の精度を評価する．

2　藤崎モデル指令列の統計的推定

藤崎モデル [1] に基づいた F_0 軌跡の統計的生成モデルと，これを用いた F_0 指令列の統計的推定 [2] の既存の枠組みを概観する．この F_0 生成モデルでは，藤崎モデルでのフレーズ・アクセント指令列 $\boldsymbol{o} = [u_\mathrm{p}[k], u_\mathrm{a}[k]]_{k=0}^{K-1}$ （k はフレームを表す添え字で，K はフレーム総数）を隠れマルコフモデル（HMM）の出力系列としてモデル化する．HMM の状態遷移図を Fig. 1 に示す．この HMM は，フレーズ指令のみ活性化した状態 p_0，アクセント指令のみ活性化した状態 $\mathrm{a}_n (n = 0, \ldots, N-1)$，いずれの指令も活性化していない状態 $\mathrm{p}_0, \mathrm{p}_1$ からなる．矩形パルスアクセント指令の持続長の分布を表現するため，状態 a_n は内部に小状態 $\mathrm{a}_{n,0}, \mathrm{a}_{n,1}, \ldots$ を持つ（Fig. 2）．一般に HMM の隠れ状態の遷移 $\boldsymbol{s} = \{s[k]\}_{k=0}^{K-1}$ の推定は EM アルゴリズムで行われる．この HMM の各状態での出力分布は正規分布で記述されており，このフレーズ指令の各フレームでの平均，アクセント指令の各状態での平均，各指令で共通の分散を定めるパラメータ群を $\theta = \{[\mu_\mathrm{p}[k]]_{k=0}^{K-1}, [\mu_{\mathrm{a},n}]_{n=0}^{N-1}, \sigma_\mathrm{p}^2, \sigma_\mathrm{a}^2\}$ とおく．また，観測 $\log F_0$ の値を $\boldsymbol{y} = [y[k]]_{k=0}^{K-1}$，藤崎モデルにおけるベースライン周波数を μ_b とおく．更に $P(\boldsymbol{y}|\boldsymbol{o}, \mu_\mathrm{b})$ として正規分布を仮定することで，F_0 軌跡の生成モデル $P(\boldsymbol{y}, \boldsymbol{o}, \boldsymbol{s}, \mu_\mathrm{b}, \theta) = P(\boldsymbol{y}|\boldsymbol{o}, \mu_\mathrm{b}) P(\boldsymbol{o}|\boldsymbol{s}, \theta) P(\boldsymbol{s}) P(\mu_\mathrm{b}) P(\theta)$ が構築される．

本モデルを用いた指令列推定では，与えられた $\boldsymbol{y}$ に対して $P(\boldsymbol{y}|\boldsymbol{o}, \theta, \mu_\mathrm{b})$ の最大化を目指して $\boldsymbol{o}, \theta, \mu_\mathrm{b}$ を反復的に更新する．これは $\boldsymbol{s}$ の分布を推定する E ステップと，E ステップの結果に基づき $\boldsymbol{o}, \theta, \mu_\mathrm{b}$ の更新

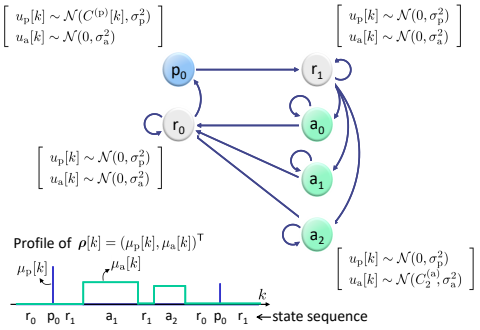

Fig. 1　F_0 軌跡の生成過程を表現する隠れマルコフモデル．1 フレームの経過毎に状態を遷移し，各状態毎に定まった確率分布に従ってフレーズ・アクセントの各指令信号を出力する．

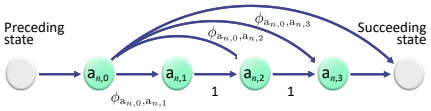

Fig. 2　各アクセント指令活性化状態 a_n を構成する内部状態のトポロジー．遷移確率の設定によってアクセント指令持続長の分布を表現する．

を行う M ステップを交互に反復する EM アルゴリズムによって行われる．

ここで，指令列を表す $\boldsymbol{o}$ は各フレーム k での状態 s_k での出力分布

$$u_\mathrm{p}[k] \sim \mathcal{N}(\mu_\mathrm{p}[k], \sigma_\mathrm{p}^2), u_\mathrm{a}[k] \sim \mathcal{N}(\mu_{\mathrm{a},s_k}, \sigma_\mathrm{a}^2)$$

に従って確率的に生成されるパラメータであり，更に確率的に推定された $\boldsymbol{s}$ の分布に従って平均された量になっている．これは単純なパルス列や矩形パルス列の形状にはなっていないため，$\boldsymbol{o}$ をそのまま推定結果とすることはできない．最終的に指令列を出力する際は事後確率最大化などの規範に基づき状態系列 $\boldsymbol{s}$ を一意に決定し，各フレームに対応する状態での出力分布の平均値 $[\mu_\mathrm{p}[k], \mu_{\mathrm{a},s_k}]_{k=0}^{K-1}$ を用いて推定結果とすることになる．この観点からは推定アルゴリズムが各反復で $\boldsymbol{s}$ に関しても最大化を行うようなものになっていることが望ましいが，以上で概観した手法は $\boldsymbol{s}$ を積分消去しているためこのような性質を満たしていない．

また，上述の手法は E ステップの数値計算に長時間を要する．これを回避するために E ステップを最

*Generative Model for Statistical Phrase/Accent Command Estimation Incorporating Phonological Features. by SATO, Ryotaro (University of Tokyo), KAMEOKA, Hirokazu, KASHINO, Kunio (NTT)

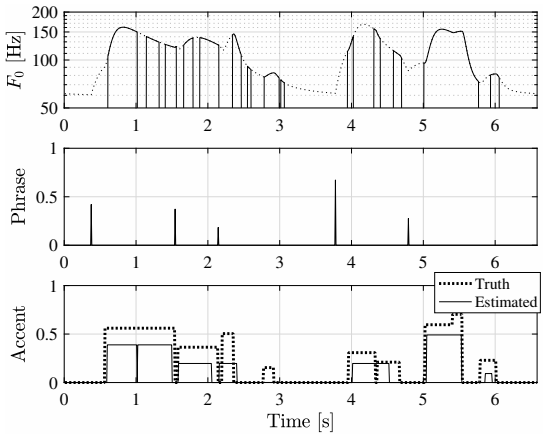

Fig. 3　音声 F_0 軌跡からのアクセント指令列推定結果の例．上 : 入力 F_0 軌跡．中 : フレーズ指令列の正解データ．下 : アクセント指令列の正解データ（点線）と推定結果（実線）．

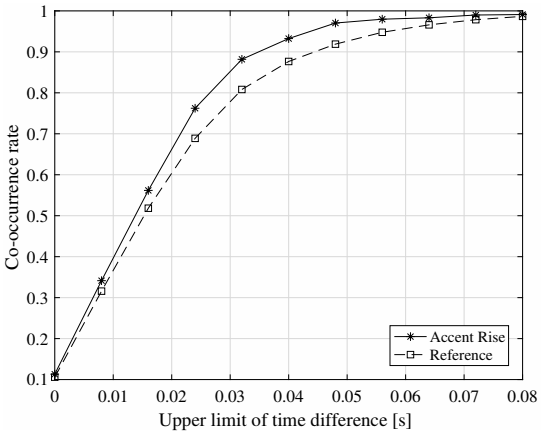

Fig. 4　実線は全アクセント指令の中で，その各立ち上がりから一定時間 S のうちに音素境界が存在するものの割合．横軸が S の値である．点線は仮にアクセント指令が有声区間に完全にランダムに存在するとした場合に同様にして算出される値．図示した範囲全てで前者が後者を上回り，アクセント指令生起時刻と音素切り替わりの間の関連性が示唆される．

尤推定に簡略化した手法が提案され，推定の精度を保ったままの大幅な高速化が実現されている [3]．この手法は，数学的には $P(\boldsymbol{y}|\boldsymbol{s},\boldsymbol{o},\theta,\mu_{\mathrm{b}})$ の $\boldsymbol{s},\boldsymbol{o},\theta,\mu_{\mathrm{b}}$ に関する最大化を行っているという解釈が可能で，先に議論したように $\boldsymbol{s}$ に関しても最大化がなされる点で先に説明した手法より優れているが，フレーズ指令の立ち上がる時刻が反復によって変化しづらいという特徴も具有する．

藤崎モデル指令列推定において，推定精度の向上は重要な課題である．アクセント指令推定精度の検討のため，指令列の正解データから再生成された F_0 軌跡を入力として，更にフレーズ指令列の初期値として正解データを用いて指令列推定を実行した結果の例を Fig. 3 に示す．この例では正解データの最初のアクセント指令を連続する 2 個の指令と誤推定しており，また逆に約 4 秒の時点から始まる 2 個のアクセント指令を誤って 1 個の指令と誤って推定している．

3　音韻特徴量時間差分を導入したモデル

アクセント指令の区切れや立ち上がりのタイミングの推定精度の向上のため，音声に含まれる F_0 軌跡以外の情報を併用して推定を行うアプローチを考える．日本語音声では，アクセント指令の出現は言語学的な高低アクセントとよく一致すると考えられている．日本語の高低アクセントはモーラ単位で定まるため，音素の区切れ位置によりアクセント指令の立ち上がりが起こりやすいような指令列の確率モデルを構築することで，より良い時間精度で，言語学的な観点からもより自然な指令列推定が行えるのではないかと期待され，このアプローチによる指令列推定の先行研究も存在する [4–6]．ATR によるコーパス音声に対して人手でラベリングされた藤崎モデル指令列について，そのアクセント指令の立ち上がりのうち，そこから一定の時間 S 以内に Julius [9] によって音素境界と認識された時刻が含まれるものの割合を示したものが Fig. 4 である．この結果からも，アクセント指令の立ち上がりが音素境界付近に現れがちであることが確かめられる．

一般に音素の変化に伴い，音韻特徴量の変動も大きくなる．したがって，音韻特徴量時間差分を F_0 指令列と同時に生成する確率モデルを構築し，アクセント指令が立ち上がるタイミングで音韻特徴量時間差分の絶対値が大きい値をとるような確率分布を導入することで，アクセント指令のタイミング推定の高精度化が達成されると期待される．

以上の仮定に基づき，[3] では，各フレームが音素境界に位置している確率を表す隠れ変数を導入し，この変数や音素境界と非音素境界における音韻特徴量時間差分の確率分布を反復毎に更新するようなアルゴリズムを提案した．この手法は，実観測 F_0 軌跡を対象とした実験でアクセント推定精度のある程度の向上を達成したものの，音韻特徴量を使用しない従来法 [2] と異なり各ステップの反復による推定値の収束が保証されないという点で理論的な面での欠点を持つ．

今回扱うモデルでは，非音素境界フレームと，音素境界に位置するフレームでの音韻特徴量時間差分の確率分布を，それぞれ別の混合ガウスモデル（Gaussian mixture model, GMM）$\mathrm{GMM}_0, \mathrm{GMM}_1$ で表現する．

また，各フレームが音素境界に位置する確率は，そのフレームでアクセント指令の立ち上がりが起きている場合は 1，そうでない場合は r とする（r は $0 \leq r \leq 1$ の範囲でチューニング可能なパラメータで，$r = 1$ で音韻特徴量を用いないモデルに帰着される）．以上により，各フレームでの HMM での隠れ状態 s と音韻特徴量時間差分 $\boldsymbol{v}$ の間に条件付き確率

$$P(\boldsymbol{v}|s) = \begin{cases} \mathrm{GMM}_1(\boldsymbol{v}) & (s = \mathrm{a}_{n,0}) \\ (1-r)\ \mathrm{GMM}_0(\boldsymbol{v}) & \\ \quad +r\ \mathrm{GMM}_1(\boldsymbol{v}) & (s \neq \mathrm{a}_{n,0}) \end{cases}$$

が導入される．

GMM は教師データを用いて事前学習しておく必要がある．音声から指令列推定を行う際は，最初にこの学習済み GMM と音韻特徴量時間差分系列データ $\boldsymbol{V} = [\boldsymbol{v}[k]]_{k=0}^{K-1}$ から確率 $P(\boldsymbol{v}[k]|s_k)$ を s_k がアクセント立ち上がりの場合とそうでない場合について計算する．音韻特徴量を用いない手法の E ステップでは $P(\boldsymbol{o}|\boldsymbol{s})P(\boldsymbol{s})$ を最大化するよう $\boldsymbol{s}$ の更新を行っていたが，本手法ではこれが $P(\boldsymbol{V}|\boldsymbol{s})P(\boldsymbol{o}|\boldsymbol{s})P(\boldsymbol{s})$ に置き換えられる．

4 実験

提案手法の性能を評価する実験を行った．ATR B セット MHT 話者の 503 文の音声データと，人手でラベリングされた各文の藤崎モデル指令列データを使用した．フレームのタイムシフトは 8 ms，音韻特徴量は MFCC（12 次）を用い，最初の 200 文の音声データを教師データとして HMM の遷移確率と GMM の学習を行い，残りの文を評価に用いた．教師データの各フレームの音素境界・非境界の分類は Julius [9] による音素セグメンテーションに基づいた．予備実験により，各 GMM の成分数は 10 とした．音声からの F_0 軌跡の抽出は中谷らの方法 [7] で行い，指令列の初期値は成澤らの方法 [8] による推定値を使用．

指令列推定の客観評価指標としては，観測された $\log F_0$ と推定された指令列から藤崎モデルの線形フィルタを通して再生成された $\log F_0$ の間の有声領域での平均二乗誤差（RMSE）の他，各指令の生起時刻のみに着目した指令脱落率・誤挿入率を用いる．これは，予め誤差として許容する時間差 S を定めておき，推定結果中の指令と正解データ中の指令の生起時刻との時間差（アクセント指令の場合は立ち上がり時刻の時間差と立ち下がり時刻の時間差の平均）が S 以内であるもの同士で重複なく最も多くのペアリングをとった際に，対応相手が存在しない指令の個数を，正解データと推定結果の各々で数え上げ，これを正解データに含まれる全指令数で割ったものとして定義

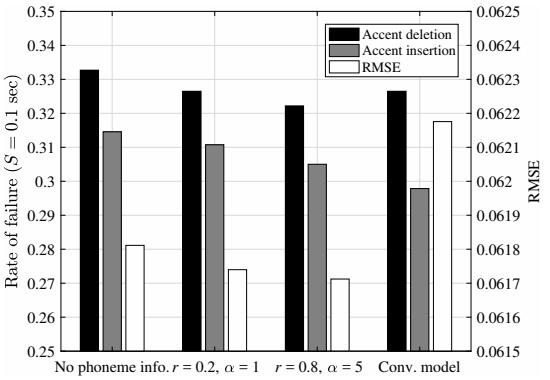

Fig. 5　観測 F_0 軌跡から指令列推定を行った結果に対する，アクセント指令の脱落率（黒色）・誤挿入率（灰色）と RMSE（白色）の比較．用いた推定法は左から順に音韻特徴量不使用のもの，本稿での提案法 2 種（$(r, \alpha) = (0.2, 1), (0.8, 5)$），[3] の手法．いずれの指標も，音韻特徴量を用いない場合と比較して改善が見られる．正解データに含まれるアクセント指令の総数は 2095 であった．

される．日本語音声のモーラ持続長はおよそ 0.1 秒程度とされるため，モーラ単位レベルでのアクセント指令の推定精度の検証を目的として今回は $S = 0.1$ sec とする．この指標は指令列の強度に関する情報を一切用いておらず，推定の妥当性は RMSE 等を併用して検証する必要がある．

指令列推定アルゴリズムとしては，音韻特徴量を全く用いないもの（E ステップは簡略化），本稿で提案したモデルで特にモデルの成り立ちから自然なパラメータをおいた（$r = 0.2$）もの，更に E ステップで最大化対象とする関数を $(P(\boldsymbol{V}|\boldsymbol{s}))^{\alpha}P(\boldsymbol{o}|\boldsymbol{s})P(\boldsymbol{s})$ とした際の重み α および r の値を最適に設定した（$r = 0.8, \alpha = 5$）もの，また [3] で提案されたものを用いた．

各手法での実験の結果得られたアクセント指令脱落・誤挿入率と RMSE を Fig. 5 に示す（フレーズ指令の脱落・誤挿入率に手法間で差異は生じなかった）．特に実観測 F_0 軌跡を入力として与えた場合，脱落・誤挿入の個数が共に最大で 3% 改善された．RMSE は音韻特徴量を使用しない手法と同程度の値となり，[3] で提案された収束が保証されない手法と比較すると 0.5%程度改善した．

具体的な音声データに対して行われた推定の例を Fig. 6 に示す．音韻情報の導入によってアクセント指令の開始が音韻特徴量の変動に伴って起きるように誘導され，この例では誤ったタイミングでのアクセント指令立ち上がりが解消されている．

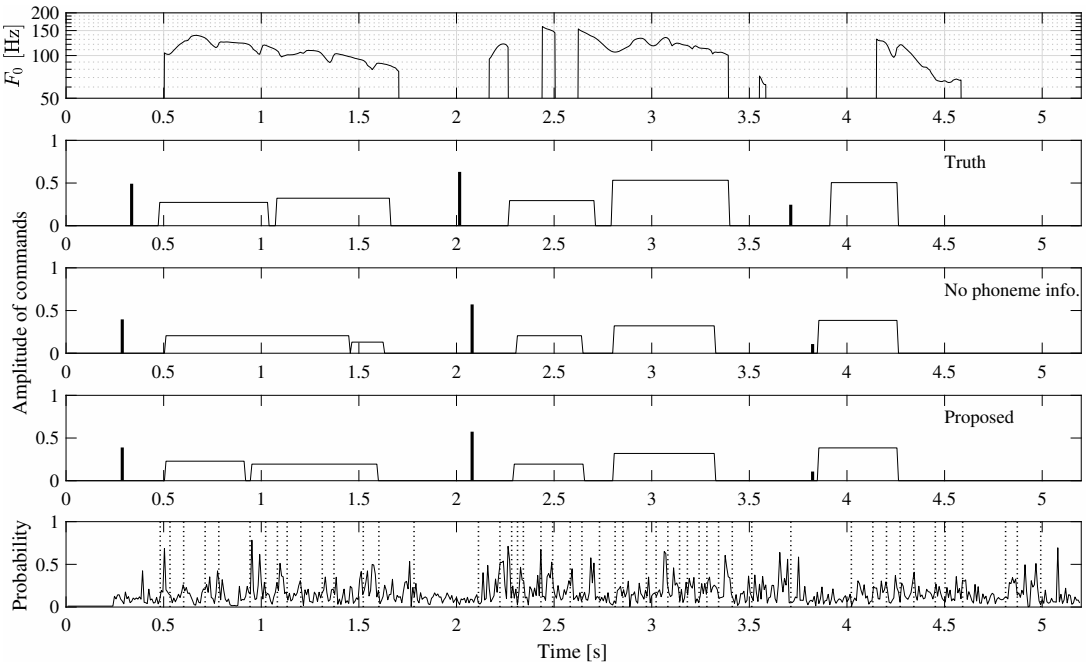

Fig. 6 「乱暴で乱暴で，ゆく先が案じられると，母がいった.」という発話文章音声の実観測 F_0 軌跡を与えた場合の指令列推定結果例．上段から観測 F_0 軌跡，正解指令列，音韻特徴量を用いない場合の推定結果，提案法（$r = 0.2, \alpha = 1$）での推定結果，各フレームが音素境界である確率．最下段の点線は Julius によって認識された音素境界時刻を示す．約 1.4 s で起きていたアクセント指令の誤った区切れは，その前後に音素境界がないことから提案法では解消されている．

5 まとめ

本稿では，藤崎モデル指令列の統計的推定問題に対して，日本語音声のアクセント指令列の出現傾向を考慮した上で，音韻情報特徴量時間差分を F_0 指令列と同時に生成する確率モデルを構成した．実音声データを用いた実験の結果，音韻特徴量を使用しないモデルと比較してアクセント指令列の推定精度のある指標における改善が確かめられた．

謝辞 本研究にあたって，ATR デジタル音声データベース B セットに対応する人手でラベル付けされた韻律パラメータのデータを提供して頂いた広瀬啓吉東京大学名誉教授に感謝いたします．

参考文献

[1] H. Fujisaki, “A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour,” Vocal Fold Physiology: Voice Production, Mechanisms and Functions, pp.347–355, 1988.

[2] H. Kameoka et al., “Generative modeling of voice fundamental frequency contours,” ACM Trans. Audio, Speech, and Language Processing, vol. 23, no. 6, pp. 1042–1053, 2015.

[3] 佐藤 他，“基本周波数パターンと音韻特徴量系列の同時生成モデルによる韻律指令列推定，” 信学技報，vol. 116, no. 378, pp.43–48, 2016.

[4] K. Hirose et al., “Use of linguistic information for automatic extraction of F_0 contour generation process model parameters.,” EUROSPEECH, pp. 141–144, 2003.

[5] H.M. Torres et al., “Linguistically motivated parameter estimation methods for a superpositional intonation model,” EURASIP Journal on Audio, Speech, and Music Processing, vol. 2014, no. 1, p. 28, 2014.

[6] K. Hirose et al., “Corpus-based extraction of F_0 contour generation process model parameters,” INTERSPEECH, pp. 3257–3260, 2005.

[7] T. Nakatani et al., “A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments,” Speech Communication, vol. 50, no. 3, pp. 203–214, 2008.

[8] S. Narusawa et al., “A method for automatic extraction of model parameters from fundamental frequency contours of speech,” Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, IEEE, pp. 509–512, 2002.

[9] http://julius.osdn.jp/