F_0 パターン生成過程の確率モデルに基づく電気音声に対するフレーズ・アクセント指令推定*

◎田中 宏（奈良先端大），亀岡 弘和（NTT），
戸田 智基（名大/奈良先端大），中村 哲（奈良先端大）

1 はじめに

喉頭摘出者のための電気式人工喉頭を用いた代替音声（電気音声）の自然性向上を目指し，我々は電気音声スペクトル特徴量系列から通常音声 F_0 パターンを予測する電気音声強調法 [1] を提案しており，自然性の大幅な向上を実験的評価により確認した．一方で，その自然性は十分に高いとは言い難く，時として物理的に発声しえない F_0 パターンが予測されうる．この問題に対して，我々は Product-of-Expert の枠組みにおいて F_0 パターンの物理的な生成過程 [2, 3] を考慮した統計的 F_0 パターン予測法 [4] を提案し，さらなる自然性の向上および F_0 予測精度の改善を実験的評価により確認した．しかしながら，[4] では，EM アルゴリズムを用いた反復的なパラメータ推定を行うため，実時間予測処理を用いたアプリケーションへの適用が困難である．

本稿では，予測時に反復処理を必要とせず，かつ，電気音声スペクトル特徴量系列に対応する物理的に生成可能な F_0 パターンを予測する手法として，F_0 パターン生成過程の確率モデルに基づき電気音声に対するフレーズ・アクセント指令推定を行う手法を提案する．評価結果より，従来法 [1] と比較して F_0 予測精度が改善されていることを示す．

2 F_0 パターン生成過程の確率モデル [2, 3]

F_0 パターンは声帯に張力を与える甲状軟骨の運動によって生み出されており，藤崎モデル [5] ではその制御機構の力学モデルが提案されている．[2, 3] では，[5] をベースにした F_0 パターン生成過程の確率モデルが提案されている．

藤崎モデル [5] では，対数 F_0 パターン $y[t]$ は，フレーズ成分 $y_p[t]$，アクセント成分 $y_a[t]$，およびベースライン成分 $y_b[t]$ の足しあわせにより表現可能と仮定されている．フレーズ成分とアクセント成分は，パルス波で表現されるフレーズ指令列 $u_p[t]$ と矩形波で表現されるアクセント指令列 $u_a[t]$ を入力とした臨界制動の二次線形系によりそれぞれ表現される．

$$y[t] = y_p[t] + y_a[t] + y_b, \quad (1)$$
$$y_p[t] = g_p[t] * u_p[t], \quad (2)$$
$$y_a[t] = g_a[t] * u_a[t]. \quad (3)$$

ここで，$*$ は畳み込みを表し，応答 $g_p[t]$ と $g_a[t]$ は話者の個人差や言語に依存しないことが経験的に知られている．

[2, 3] では，藤崎モデルの両指令列 $u_p[t]$ および $u_a[t]$ を以下に示す経路制約付き HMM の出力系列と見なすことにより，$P(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F)$ を記述する．ここで，$\boldsymbol{u} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_T]^\top$ であり，$\boldsymbol{\lambda}_F$ は HMM のパラメータセットを表す．

出力系列: $\boldsymbol{u}_t = (u_p[t], u_a[t])^\top \quad (t = 1, \cdots, T)$
状態集合: $\boldsymbol{S} = \{r_0, p_0, \cdots, p_{M-1}, r_1, a_0, \cdots, a_{N-1}\}$
状態系列: $\boldsymbol{s} = \{s_t \in \boldsymbol{S} | t = 1, \cdots, T\}$
出力分布: $P(\boldsymbol{u}_t | s_t = i) = \mathcal{N}(\boldsymbol{u}_t; \boldsymbol{c}_i, \boldsymbol{\Sigma}_i)$

$$\boldsymbol{c}_i = \begin{cases} (0,0)^\top & (i \in r_0, r_1) \\ (\mu_p^{(m)}, 0)^\top & (i \in p_m) \\ (0, \mu_a^{(n)})^\top & (i \in a_n) \end{cases} \quad \boldsymbol{\Sigma}_i = \begin{bmatrix} v_{p,i}^2 & 0 \\ 0 & v_{a,i}^2 \end{bmatrix}.$$

遷移確率: $\phi_{\hat{i},i} = P(s_t = i | s_{t-1} = \hat{i})$

また，式 (1)-(3) より $P(\boldsymbol{y}|\boldsymbol{u})$ を定義することができ，藤崎モデルに準拠した $\boldsymbol{y} = [y[1], \cdots, y[T]]^\top$ の生成モデルを以下のように導くことができる．

$$p(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\lambda}_F) = \int p(\boldsymbol{y}|\boldsymbol{u}) p(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F) \, \mathrm{d}\boldsymbol{u} \quad (4)$$
$$= \mathcal{N}(\boldsymbol{y}; \boldsymbol{m}, \boldsymbol{\Gamma}). \quad (5)$$

ここで，$\boldsymbol{m} = [m[1], \cdots, m[T]]^\top$ は与えられた状態系列 $\boldsymbol{s}$ から決定される藤崎モデルに準拠した F_0 パターンであり，$m[t] = g_p[t] * \mu_p[t] + g_a[t] * \mu_a[t] + y_b$ である．モデルパラメータを適切に学習することにより，与えられた F_0 パターンに対して統計的に尤度の高いフレーズ指令列およびアクセント指令列を推定可能となる．

3 電気音声強調のための F_0 パターン生成過程の確率モデル

電気音声に対して，F_0 パターン生成過程の物理的制約を満たす自然な F_0 パターンを予測する．本提案法では，電気音声のスペクトル特徴量系列とそれに対応する通常音声のフレーズ指令列およびアクセント指令列との結合確率がモデル化される．2 節で述べた状態系列 $\boldsymbol{s}$ は通常音声のフレーズ指令列およびアクセント指令列に関連付いているため，各状態において電気音声のスペクトル特徴量の出力分布を GMM で

*Estimation of phrase and accent commands from electrolaryngeal speech using probabilistic model of F_0 generative process. by TANAKA, Kou (NAIST), KAMEOKA, Hirokazu (NTT), TODA, Tomoki (Nagoya University/NAIST), and NAKAMURA, Satoshi (NAIST)

表現することにより，HMM-GMM の出力系列および状態系列の結合確率は，電気音声から通常音声のフレーズ指令列およびアクセント指令列を予測するモデルとして使用可能である．最終的に得られる F_0 パターンは，入力された電気音声のスペクトル特徴量系列に対して統計的に尤度の高いフレーズ指令列およびアクセント指令列を予測したのち，予測された両司令列を入力とする臨界制動の二次線形系の足し合わせで表現される．

学習処理では，電気音声と通常音声のパラレルデータを用いて，電気音声のスペクトル特徴量系列 $\boldsymbol{X}$ と通常音声のフレーズ指令列およびアクセント指令列の状態系列 $\boldsymbol{s}$ の結合確率が新たにモデル化される．

$$p(\boldsymbol{X}, \boldsymbol{s}|\boldsymbol{\lambda}_F) = p(\boldsymbol{X}|\boldsymbol{s}, \boldsymbol{\lambda}_F)p(\boldsymbol{s}|\boldsymbol{\lambda}_F). \quad (6)$$

ここで，

$$p(\boldsymbol{s}|\boldsymbol{\lambda}_F) = \prod_t p(s_t|s_{t-1}, \boldsymbol{\lambda}_F), \quad (7)$$

$$p(\boldsymbol{X}|\boldsymbol{s}, \boldsymbol{\lambda}_F) = \prod_t p(\boldsymbol{X}_t|\boldsymbol{s}_t, \boldsymbol{\lambda}_F), \quad (8)$$

であり，各状態における電気音声のスペクトル特徴量に関する出力分布は以下の GMM で記述される．

$$p(\boldsymbol{X}_t|s_t = i) = \sum_m \alpha_{i,m}\mathcal{N}(\boldsymbol{X}_t; \boldsymbol{\mu}_{i,m}, \boldsymbol{\Sigma}_{i,m}). \quad (9)$$

ここで，$\alpha_{i,m}$，$\boldsymbol{\mu}_{i,m}$ および $\boldsymbol{\Sigma}_{i,m}$ は状態 i における m 番目の正規分布の混合重み，平均ベクトルおよび分散行列である．

予測処理では，電気音声のスペクトル特徴量が与えられたもとで，以下の条件つき確率を最大化する状態系列 $\hat{\boldsymbol{s}}$ を予測する．

$$\hat{\boldsymbol{s}} = \underset{\boldsymbol{s}}{\operatorname{argmax}}\, p(\boldsymbol{s}|\boldsymbol{X}, \boldsymbol{\lambda}_F) = \underset{\boldsymbol{s}}{\operatorname{argmax}}\, p(\boldsymbol{X}, \boldsymbol{s}|\boldsymbol{\lambda}_F). \quad (10)$$

2 節で述べた通り，一度 $\boldsymbol{s}$ が決まると，式 (4) を最大化する F_0 パターン $\boldsymbol{y}$ は解析的に導出可能である．なお，EM アルゴリズムを用いた反復的なパラメータ推定 [4] と比較して，Viterbi アルゴリズムに基づく状態系列 $\hat{\boldsymbol{s}}$ の予測は，実時間予測処理への適用が容易である．

4 実験的評価

4.1 実験条件

入力音声は男性健常者 1 名による模擬電気音声，目標音声は異なる男性健常者 1 名による通常音声である．学習データとして ATR 音素バランス文 503 文中 450 文を用い，評価データとして残り 50 文を用いる．フレームシフト長は 5 ms とする．F_0 パターン生成過程の確率モデルに関するその他の学習条件は [3] と同条件とする．

客観評価として，以下のシステムにおける予測 F_0 パターンと目標音声の F_0 パターンとの間の相関係数を計算する．

Table 1　F_0 相関係数に関する客観評価結果.

gmm	0.40
gmm + post	0.41
proposed	0.50

gmm ：GMM に基づく統計的 F_0 予測法 [1].

gmm + post ： [1] で予測された F_0 パターン ***gmm*** に対して， 2 節の方法で藤崎モデルを適用することで F_0 パターンを修正する手法.

proposed ：提案法.

4.2 実装評価結果

表 1 に各システムにおける予測 F_0 パターンと目標音声の F_0 パターンとの間の相関係数を示す．***gmm*** と比較すると，***gmm*+*post*** ではわずかに，***proposed*** では大幅に改善していることがわかる．これより，電気音声強調において F_0 パターン生成過程の制約を取り入れることは有効である．また，***proposed*** は，***gmm + post*** よりも高い値をとることから，F_0 パターン生成過程の制約を後処理として導入する枠組みでは，十分な効果が得られないことが分かる．これに対し，提案法では，制約を組み込んだ予測処理を可能とすることで，予測精度の改善が得られる．

5 おわりに

本稿では，電気音声強調において予測される F_0 パターンのさらなる自然性改善のため，電気音声のスペクトル特徴量系列から通常音声のフレーズ指令列およびアクセント指令列を F_0 パターン生成過程の確率モデルに基づき推定した．実験的評価において，F_0 パターン予測精度の向上を確認した．

謝辞　本研究の一部は，JSPS 科研費 15J10727 および 26280060 の助成を受け実施したものである．

参考文献

[1] K. Tanaka *et al.*, *IEICE Trans. Information and Systems,* Vol. E97-D, No. 6, pp. 1429–1437, Jan. 2014.

[2] K. Yoshizato *et al.*, *Proc. Speech Prosody,* pp. 175–178, May 2012.

[3] H. Kameoka *et al.*, *IEEE/ACM Trans. Audio, Speech, and Language,* Vol. 23, No. 6, pp. 1042–1053, Jun. 2015.

[4] K. Tanaka *et al.*, *Proc. ICASSP,* pp. 5665–5669, Mar. 2016.

[5] H. Fujisaki, *Vocal Fold Physiology: Voice Production, Mechanisms and Functions,* pp. 347–355, 1998.