F_0 パターン生成過程を考慮した Product-of-Experts に基づく電気音声強調のための統計的 F_0 予測法 *

◎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）の自然性を向上することを目的とし，我々は統計的 F_0 予測法に基づく電気音声強調法 [1] を提案しており，実験的評価により自然性の向上を確認している．しかしながら，この手法で予測される F_0 パターンの自然性は十分に高いとは言い難く，時には物理的に人間が発声し得ないような不自然な F_0 パターンが予測され得る．この問題に対し，F_0 パターンの物理的な生成過程を考慮した予測を行うことで，より自然な F_0 パターンを生成できる可能性がある．

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，[2, 3] ではその制御機構の確率モデルが提案されている．本稿では，[2, 3] の確率モデルと [1] の確率モデルを Product-of-Experts(PoE) の枠組により統合することで，F_0 パターンの物理的な生成過程の制約を考慮した上で，電気音声スペクトル特徴量系列に対応する F_0 パターンの予測を可能とする統計的 F_0 予測法を提案する．実験的評価より，提案法により，F_0 予測精度および強調音声の自然性が改善されることを示す．

2 電気音声のための統計的 F_0 予測法 [1]

[1] では，統計的手法に基づき，電気音声のスペクトル系列から通常音声の F_0 パターンを予測する手法が提案されている．本手法は，学習処理と変換処理で構成される．

学習処理では，電気音声と通常音声の同一発話データを用いて，変換モデルが学習される．各時間フレームにおいて，前後数フレームから得られる電気音声のスペクトルセグメント特徴量と，通常音声の対数 F_0 に対する静的・動的特徴量を抽出したのち，スペクトル距離尺度に基づく動的時間伸縮によりこれらを対応付けた結合ベクトルを用いて，両特徴量間の結合確率密度関数が GMM でモデル化される．

変換処理では，学習された GMM を用いて，次式で示される最尤系列変換法 [5] により，電気音声のスペクトルセグメント特徴量系列 $\boldsymbol{x}$ は通常音声の対数 F_0 系列 $\boldsymbol{y}$ へと変換される．

$$\hat{\boldsymbol{y}} = \mathop{\mathrm{argmax}}_{\boldsymbol{y}} P(\boldsymbol{o}|\boldsymbol{x}, \boldsymbol{\lambda}_G) \quad \text{subject to } \boldsymbol{o} = \boldsymbol{W}\boldsymbol{y}$$

$$\begin{aligned} P(\boldsymbol{o}|\boldsymbol{x}, \boldsymbol{\lambda}_G) &= \sum_{\boldsymbol{m}} P(\boldsymbol{o}|\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{\lambda}_G) P(\boldsymbol{m}|\boldsymbol{x}, \boldsymbol{\lambda}_G) \\ &\simeq P(\boldsymbol{o}|\boldsymbol{x}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}_G) P(\hat{\boldsymbol{m}}|\boldsymbol{x}, \boldsymbol{\lambda}_G) \end{aligned}$$

ここで，$\boldsymbol{\lambda}_G$ は GMM のパラメータセットを表し，$\boldsymbol{W}$ は静的特徴量系列 $\boldsymbol{y}$ を静的・動的特徴量系列 $\boldsymbol{o}$ に写像する変換行列を表す．また，$\boldsymbol{m}$ は分布系列を，$\hat{\boldsymbol{m}}$ は準最適な分布系列を表す．

3 F_0 パターン生成過程の確率モデル [2, 3]

[2, 3] では，藤崎モデル [4] をベースにした F_0 パターンの生成過程の確率モデルが提案されている．以下でその概要を述べる．

藤崎モデルは，フレーズ成分 $y_{p(t)}$，アクセント成分 $y_{a(t)}$（t は時刻），およびベースライン成分 y_b の足しあわせにより対数 F_0 パターン $y_{(t)}$ を表現可能と仮定するモデルである．フレーズ成分とアクセント成分は，フレーズ指令と呼ばれるパルス波の列 $u_{p(t)}$ とアクセント指令と呼ばれる矩形波の列 $u_{a(t)}$ を入力とした臨界制動の二次線形系により表現される．

$$\begin{aligned} y_{(t)} &= y_{p(t)} + y_{a(t)} + y_b \\ y_{p(t)} &= g_{p(t)} * u_{p(t)} \\ y_{a(t)} &= g_{a(t)} * u_{a(t)} \end{aligned}$$

ここで，$*$ は畳み込みを表す．また，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．

藤崎モデルの両指令列 $u_{p(t)}$ および $u_{a(t)}$ を以下に示すトポロジー（経路制約付き HMM）の出力系列と見なすことにより，$P(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F)$ を記述することができる．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: $P(\boldsymbol{u}_{(t)}|s_t = i) = \mathcal{N}(\boldsymbol{u}_{(t)}; \boldsymbol{c}_i, \boldsymbol{\Sigma}_i)$

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

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

また，$y_{(t)} = g_{p(t)} * u_{p(t)} + g_{a(t)} * u_{a(t)} + y_b$ という関係式より $P(\boldsymbol{y}|\boldsymbol{u})$ を定義することができ，$P(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F) = P(\boldsymbol{y}|\boldsymbol{u}) p(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F)$ を $\boldsymbol{u}$ について周辺化することで藤崎モデルに準拠した $\boldsymbol{y}$ の生成モデル $P(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\lambda}_F)$ を導くことができる．

4 PoE に基づく統計的 F_0 予測法

PoE は観測データの確率分布を Experts と呼ばれる個々の確率分布の積で表現する手法であり，そのモデルは複数の Expert の論理積（AND）を取ったような確率モデルとなる．本研究では，2 節で述べた統計的 F_0 予測法 [1] における GMM と 3 節で述べた F_0 パターン生成過程の確率モデル [2, 3] を PoE として統合することで，F_0 パターンの物理的な生成過程の制約の下で，電気音声のスペクトル系列に対応する F_0 パターンを統計的に予測する手法を提案する．なお，PoE による確率モデルは，正規化項の計算が困難となるため，モデルパラメータの推定には，サンプリングに基づく近似がよく用いられる．これに対し，本研究では，2 節で述べた GMM を潜在系列モデル [6] へと定義し直すことで，EM アルゴリズムを用いたパラメータ推定を可能とする手法を提案する．

4.1 潜在系列モデル

2 節で述べたモデルにおいて，静的・動的特徴量系列 $\boldsymbol{o}$ を静的特徴量系列 $\boldsymbol{y}$ の関数 $\boldsymbol{o} = \boldsymbol{W}\boldsymbol{y}$ として扱う代わりに，$\boldsymbol{o}$ を潜在変数とみなし，緩い制約 $\boldsymbol{o} \simeq \boldsymbol{W}\boldsymbol{y}$ の下で $\boldsymbol{y}$ が生成されると仮定すると，完全

* Statistical F_0 Prediction For Electrolaryngeal Speech Enhancement Based on Product-of-Experts Framework Considering F_0 Generative Process. by TANAKA, Kou (NAIST), KAMEOKA, Hirokazu (NTT), TODA, Tomoki (Nagoya University/NAIST), and NAKAMURA, Satoshi (NAIST)

データ $\boldsymbol{y},\boldsymbol{x},\boldsymbol{o},\boldsymbol{m}$ に対する確率密度関数は以下のように定義される．

$$P(\boldsymbol{y},\boldsymbol{x},\boldsymbol{o},\boldsymbol{m}|\boldsymbol{\lambda}_G) = P(\boldsymbol{y}|\boldsymbol{o})P(\boldsymbol{x},\boldsymbol{o}|\boldsymbol{m},\boldsymbol{\lambda}_G)P(\boldsymbol{m}|\boldsymbol{\lambda}_G) \quad (1)$$

一方で，3 節で述べたモデルにおいては，完全データ $\boldsymbol{y},\boldsymbol{u},\boldsymbol{s}$ に対する確率密度関数は以下のように定義される．

$$P(\boldsymbol{y},\boldsymbol{u},\boldsymbol{s}|\boldsymbol{\lambda}_F) = P(\boldsymbol{y}|\boldsymbol{u})P(\boldsymbol{u}|\boldsymbol{s},\boldsymbol{\lambda}_F)P(\boldsymbol{s}|\boldsymbol{\lambda}_F), \quad (2)$$

どちらのモデルも，観測データ系列は，潜在変数系列を通して生成される点に注意する．

4.2 PoE

式（1）および式（2）で表される潜在系列モデルを PoE の枠組みで統合する．この時，観測データ系列 $\boldsymbol{y}$ と GMM の分布系列 $\boldsymbol{m}$ および HMM の状態系列 $\boldsymbol{s}$ の間には，潜在変数系列 $\boldsymbol{o}$ および $\boldsymbol{u}$ が与えられた下での条件付き独立性が成り立つため，PoE モデルの確率密度関数を解析的に求めることが容易となる．完全データ $\boldsymbol{y},\boldsymbol{x},\boldsymbol{o},\boldsymbol{u},\boldsymbol{m},\boldsymbol{s}$ に対する確率密度関数は以下のようになる．

$$\begin{aligned}
&P(\boldsymbol{y},\boldsymbol{x},\boldsymbol{o},\boldsymbol{u},\boldsymbol{m},\boldsymbol{s}|\boldsymbol{\lambda}_G,\boldsymbol{\lambda}_F)\\
&= \underbrace{P(\boldsymbol{y}|\boldsymbol{o},\boldsymbol{u})P(\boldsymbol{x},\boldsymbol{o}|\boldsymbol{m},\boldsymbol{\lambda}_G)P(\boldsymbol{u}|\boldsymbol{s},\boldsymbol{\lambda}_F)}_{P(\boldsymbol{y},\boldsymbol{x},\boldsymbol{o},\boldsymbol{u}|\boldsymbol{m},\boldsymbol{s},\boldsymbol{\lambda}_G,\boldsymbol{\lambda}_F)} P(\boldsymbol{m}|\boldsymbol{\lambda}_G)P(\boldsymbol{s}|\boldsymbol{\lambda}_F)\\
&P(\boldsymbol{y}|\boldsymbol{o},\boldsymbol{u}) \propto \mathcal{N}(\boldsymbol{y};\boldsymbol{\mu}_{y|o},\boldsymbol{V})\cdot\mathcal{N}(\boldsymbol{y};\boldsymbol{\mu}_{y|u},\boldsymbol{\Gamma})\\
&\qquad = \mathcal{N}(\boldsymbol{y};\boldsymbol{\mu}_{y|o,u},\boldsymbol{\Sigma}_{y|o,u}),\\
&\boldsymbol{\mu}_{y|o,u} = (\boldsymbol{V}^{-1}+\boldsymbol{\Gamma}^{-1})^{-1}(\boldsymbol{V}^{-1}\boldsymbol{\mu}_{y|o}+\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}_{y|u}), \quad (3)\\
&\boldsymbol{\Sigma}_{y|o,u} = (\boldsymbol{V}^{-1}+\boldsymbol{\Gamma}^{-1})^{-1},
\end{aligned}$$

ここで，$\boldsymbol{V}$ および $\boldsymbol{\Gamma}$ は正定値行列である．この時，以下に示す Q 関数を設定することで，EM アルゴリズムを用いたパラメータ（F_0 パターン $\boldsymbol{y}$，GMM の分布系列 $\boldsymbol{m}$，HMM の状態系列 $\boldsymbol{s}$）の推定が可能となる．

$$\begin{aligned}
Q(\theta,\theta') =& \mathbb{E}_{\boldsymbol{o},\boldsymbol{u}|\tilde{\boldsymbol{x}},\boldsymbol{y}',\boldsymbol{m}',\boldsymbol{s}'}\left[\log P(\boldsymbol{y},\tilde{\boldsymbol{x}},\boldsymbol{o},\boldsymbol{u}|\boldsymbol{m},\boldsymbol{s},\hat{\boldsymbol{\lambda}}_G,\hat{\boldsymbol{\lambda}}_F)\right]\\
&+\log P(\boldsymbol{m}|\hat{\boldsymbol{\lambda}}_G)+\log P(\boldsymbol{s}|\hat{\boldsymbol{\lambda}}_F),\\
&\{\boldsymbol{y},\boldsymbol{m},\boldsymbol{s}\} \leftarrow \underset{\boldsymbol{y},\boldsymbol{m},\boldsymbol{s}}{\operatorname{argmax}} Q(\theta,\theta').
\end{aligned}$$

5 実験的評価

5.1 実験条件

入力音声は男性喉頭摘出者 1 名による電気音声，目標音声は女性健常者 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 A セットの 50 文中 40 文を用い，評価データとして残り 10 文を用いた 5 交差検定を行う．フレームシフト長は 5 ms とする．GMM に基づく統計的 F_0 予測法に関するその他の学習条件は [1] と同条件，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．主観評価として，音声の自然性に関する 5 段階オピニオン評定を行う．被験者は男性 5 名である．

- *GMM-based*: GMM に基づく統計的 F_0 予測法．
- *Fujisaki-model-based1*: *GMM-based* で予測された F_0 パターンに対して，3 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．ただし，E ステップを簡易化し，予測された F_0 パターンを新たな観測データとみなして，GMM の分布系列の更新と HMM の状態系列の更新を実装したもの．
- *Fujisaki-model-based2*: *Proposed* で予測された F_0 パターンに対して，3 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．

Table 1　F_0 相関係数に関する客観評価結果．

GMM -based	Fujisaki -model -based1	Proposed	Fujisaki -model -based2
0.56	0.49	0.59	0.56

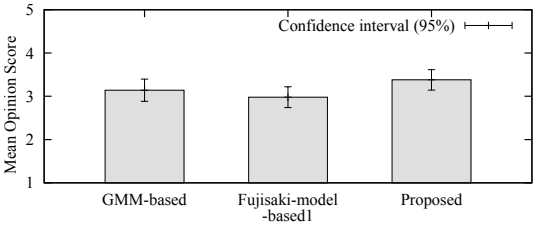

Fig. 1　自然性に関する主観評価結果．

5.2 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．*Proposed* が最も高い値をとることから，提案法である PoE の有効性が確認できる．また，*Fujisaki-model-based1* および *Fujisaki-model-based2* は，各々 *GMM-based* および *Proposed* よりも低い値をとることから，電気音声強調のように予測 F_0 パターンの精度が限定的である場合，[7] で報告されているような 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．

図 1 に自然性に関する主観評価結果を示す．*GMM-based* や *Fujisaki-model-based1* よりも *Proposed* が高い自然性を有することがわかる．

6 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，Product-of-Experts の枠組みにおいて F_0 生成過程モデルを考慮した統計的 F_0 予測法を提案した．実験的評価より，F_0 予測精度および音声の自然性が改善することを示した．なお，本稿では，提案法における E ステップに対して近似処理を導入した．近似処理を無くすことでさらなる性能改善が期待されるため，今後，提案法の厳密な評価を行う．

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] H. Fujisaki, *Vocal Fold Physiology: Voice Production, Mechanisms and Functions,* pp. 347–355, 1998.
[5] T. Toda *et al.*, *IEEE Trans. Audio, Speech, and Language,* Vol. 15, No. 8, pp. 2222–2235, Nov 2007.
[6] H. Kameoka, *Proc. The 25th IEEE International Workshop. Machine Learning for Signal Processing,* Sep. 2015.
[7] T. Matsuda *et al.*, *Acoustical Science and Technology,* Vol. 33, No. 4, pp. 221–228, 2012.