VAE-SPACE: 音声 F_0 パターンの深層生成モデル*

◎田中 宏，亀岡 弘和（NTT)，森川 一穂（名大)

1 はじめに

基本周波数（F_0）パターンは，言語・非言語情報と深い関係がある．例えば，話者は，発話文の語尾の F_0 パターンを変化させることで疑問文を表現し，F_0 パターンのダイナミクスを変化させることで意図や感情を表現する．また，歌声においても，メロディや情感，個人性を表現するために，歌唱者は F_0 パターンを変化させる．F_0 パターン生成過程のモデル化は，表現豊かな音声・歌声合成や対話システム，話者・感情認識などの実現に極めて有効である．

話し声の F_0 パターン生成過程のモデルとしては，喉頭による声帯の制御機構を模擬した物理モデル（通称「藤崎モデル」）[1] とそのパラメータを統計的手法により推定することを可能にする藤崎モデルの確率モデル版 [2] が提案されている．藤崎モデルは歌声には当てはまらない仮定をいくつか置くため，そのままの形では歌声の F_0 パターンに適用することはできないが，藤崎モデルを歌声の特徴に合わせて適応したいわば藤崎モデルの歌声版も提案されている [3, 4]．これらのモデルでは，F_0 パターンを直感的かつ解釈可能な生成過程パラメータ（藤崎モデルではフレーズ成分やアクセント成分，歌声モデルでは楽譜情報またはメロディ成分や表現成分に相当）により記述し，所与の F_0 パターンからこれらを推定することで話し声や歌声の特徴を保持したまま自由に F_0 パターンを加工したり変換したりすることを可能とする．しかし，これらのモデルで共通する問題として，それぞれ話し声（特定の発話スタイルや言語）や歌声（特定の歌唱スタイル）に特化したモデルとなっている点とパラメータ推定のために計算コストの高い反復アルゴリズムを要する点が挙げられ，これらが用途を限定的にしている．

本稿では，深層生成モデルに基づき，音声・歌声に特化しない F_0 パターンの普遍的な生成過程モデルとその内部パラメータを高速かつ高精度に推定するアルゴリズムとを学習により同時に発見することを可能にする方法論を提案する．

2 F_0 パターン生成過程モデル

2.1 音声の F_0 パターン生成過程モデル

音声の F_0 パターンは，韻律句全体にわたってゆるやかに変化する成分 (フレーズ成分) と，アクセントに従って急峻に変化する成分 (アクセント成分) により構成される．藤崎モデル [1] は，甲状軟骨の運動による F_0 パターンの生成過程を説明した物理モデルであり，フレース成分 $y_{p(t)}$，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 を用いて対数 F_0 パターン $y_{(t)}$ を $y(t) = x_{\mathrm{p}}(t) + x_{\mathrm{a}}(t) + \mu_b$ と表現するモデルである．

2.2 歌声の F_0 パターン制御モデル

歌声の F_0 パターンは，歌のメロディ成分と，急激な変化（オーバーシュートなど）や周期的な変化（ビブラート）などの混合成分により構成される．音声の F_0 パターンと比較すると，オーバーシュートやビブラートを含む歌声の F_0 パターンは，上述した藤崎モデルでは単純に表現できない．そのため，歌声の F_0 パターン制御モデル [4] では，制御パラメータ（減衰率 ζ と固有周波数 Ω）を用いて表現される 2 次系の伝達関数 $\mathcal{G}(s) = \Omega^2/(s^2 + 2\zeta\Omega s + \Omega^2)$ における減衰率 ζ を調整することによって，指数減衰（$\zeta > 1$），減衰振動（$0 < \zeta < 1$，オーバーシュートに対応する），臨界制動（$\zeta = 1$），定常振動（$\zeta = 0$，ビブラートに対応する）からなる様々な振動現象を表現する．

3 VAE-SPACE: 音声 F_0 パターンの深層生成モデル

3.1 コンセプト

深層ニューラルネットワーク (Deep Neural Network; NN) を用いた生成モデルを深層生成モデルといい，その一種である変分自己符号化器 (Variational Autoencoder; VAE) [5, 6] が近年画像生成や音声変換などのタスクにおいて高い効果を示している．VAE は，その名称が示すように，自己符号化器 (Autoencoder; AE) の確率モデル版であり，入力データ（例えば，音声や画像）が与えられたもとでの潜在変数の条件付分布のパラメータを出力するエンコーダ NN と，潜在変数が与えられたもとでの入力データの条件付分布のパラメータを出力するデコーダ NN からなる． 従来の VAE では潜在変数には特定の事前分布（正規分布など）に従うこと以外の仮定は特に置かれないが，所望の仮定を確率モデルの形でより詳しく記述し，潜在変数の事前分布として導入することができれば，デコーダは観測データとその観測データに内在する解釈可能なパラメータとを関連づける強力な生成モデル（例えば，F_0 パターンとフレーズ・アクセント成分とを関連づける藤崎モデルのような生成モデル）になりうる．さらに，デコーダとともにエンコーダを学習することで，観測データの生成過程モデルとともに観測データから所望の生成過程パラメータを推論するアルゴリズムに相当する NN を同時に得ることができる．また，観測データは大量にある一方で観測データと生成過程パラメータのペアデータの量が限定的な状況でも，VAE の生成モデルとしての性質を活かして半教師あり学習を行える点も大きな魅力である．

そこで本稿では，VAE に基づき，F_0 パターンの生成過程モデルとその内部パラメータを推定するアルゴリズムとを学習により発見することを可能にする VAE-SPACE(Statistical Phrase/Accent Command Estimation) を提案する．なお，本モデルでは F_0 パターン全体を観測データとする．また，エンコーダとデコーダは再帰型 NN を用いてモデル化することもできるが，並列計算に向いた畳み込み NN(Convolutional NN; CNN) により記述することでフォワードパスの計算を高速に実現することができる．

---

*VAE-SPACE: Deep Generative Model for Voiced F_0 contours. by TANAKA, Kou (NTT), KAMEOKA, Hirokazu (NTT), MORIKAWA, Kazuho (Nagoya University)

Table 1　歌声 F_0 パターン類似性に関する聴取実験結果 (42 サンプル ×7 評価者).

VAE-SPACE	Musical score or Fair	p-value
76.2 %	23.8（Fair: 16.7） %	0.000312

Table 2　音声 F_0 パターンにおける推定誤差（53 文平均）.

	VAE-SPACE	SPACE
F_0 contour	**0.0536**	0.0883
Phrase component	**0.0947**	0.123
Accent component	**0.0936**	0.122

3.2　VAE-SPACE

潜在変数 $\boldsymbol{z}$ を，F_0 パターンの生成過程を司るパラメータ（例えば，藤崎モデルの場合ではフレーズ･アクセント成分）とする．$\boldsymbol{z}$ を入力とし，F_0 パターン $\boldsymbol{x}$ の条件付分布 $P_\theta(\boldsymbol{x}|\boldsymbol{z})$ のパラメータを出力するデコーダ NN は F_0 パターン生成過程モデルと見なせる．一方，$\boldsymbol{z}$ の事後分布 $P_\theta(\boldsymbol{z}|\boldsymbol{x})$ は所与の F_0 パターン $\boldsymbol{x}$ から $\boldsymbol{z}$ を推論する逆過程と見なせる．この事後分布を $P_\theta(\boldsymbol{x}|\boldsymbol{z})$ から厳密に算出することは難しいが，代わりに分布 $Q_\phi(\boldsymbol{z}|\boldsymbol{x})$ のパラメータを出力するエンコーダ NN を別に導入し，$Q_\phi(\boldsymbol{z}|\boldsymbol{x})$ が $P_\theta(\boldsymbol{z}|\boldsymbol{x}) \propto P_\theta(\boldsymbol{x}|\boldsymbol{z})P(\boldsymbol{z})$ にできるだけ近くなるようにデコーダ NN とエンコーダ NN を学習することができる．F_0 パターン $\boldsymbol{x}$ に関する対数周辺確率密度関数 $\log P_\theta(\boldsymbol{x})$ は，

$$\log P_\theta(\boldsymbol{x}) = \mathcal{L}(\theta,\phi;\boldsymbol{x}) + D_{\mathrm{KL}}\left[Q_\phi(\boldsymbol{z}|\boldsymbol{x})||P_\theta(\boldsymbol{z}|\boldsymbol{x})\right], \quad (1)$$

$$\mathcal{L}(\theta,\phi;\boldsymbol{x}) = -D_{\mathrm{KL}}\left[Q_\phi(\boldsymbol{z}|\boldsymbol{x})||P(\boldsymbol{z})\right] + \mathbb{E}_{Q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log P_\theta(\boldsymbol{x}|\boldsymbol{z})\right] \quad (2)$$

で与えられる．ただし，ここで $D_{\mathrm{KL}}[\cdot|\cdot]$ は Kullback-Leibler（KL）距離を表す．式 (2) より，θ と ϕ について $\mathcal{L}(\theta,\phi;\boldsymbol{x})$ を最大化することが，$P_\theta(\boldsymbol{z}|\boldsymbol{x})$ と $Q_\phi(\boldsymbol{z}|\boldsymbol{x})$ との KL 距離を最小化することに相当する．ここでは，$Q_\phi(\boldsymbol{z}|\boldsymbol{x})$ と $P_\theta(\boldsymbol{x}|\boldsymbol{z})$ の分布形は正規分布とする．ここで，潜在変数 $\boldsymbol{z}$ に満たしてほしい特定の仮定があれば，事前分布 $P(\boldsymbol{z})$ を通して設定可能である．例えば，上述のように潜在変数 $\boldsymbol{z}$ をフレーズ･アクセント成分と関連づける場合，$P(\boldsymbol{z})$ は $P(\boldsymbol{z}) = \sum_{\boldsymbol{s}} P(\boldsymbol{z}|\boldsymbol{s})P(\boldsymbol{s})$ とすることができる．なお、$\boldsymbol{s}$ は，[2] で述べられている経路制約付き HMM の状態系列である．

4　実験的評価

4.1　実験条件

音声データセット（学習データ：429 文［30 分］，評価データ：53 文［3 分］）と歌声データセット（学習データ：42 セット［15 分］，評価方法：Leave-one-out 交差検証）を用いて，評価実験を行った．従来法 (**SPACE**) に関するモデル条件は，[2] と同等とした．提案法 (**VAE-SPACE**) のモデル構造は，5 層の CNN により記述されたエンコーダおよび 2 層の CNN により記述されたデコーダとした．

Table 3　F_0 パターンから藤崎モデルのフレーズおよびアクセント成分を推定するのに要する処理時間（単位: 秒）.

	VAE-SPACE	SPACE
53 sentences	**0.0126** ± 0.0002	2712.080
average	—	5.67

4.2　歌声の F_0 パターン類似性に関する聴取実験（**XAB** テスト）

提案手法における潜在変数を MIDI を表現する変数と仮定し学習したのち，デコーダへ MIDI を入力し F_0 パターンを推定した．表 1 より，提案手法が目標とする F_0 パターンを再現できていることがわかる．

4.3　音声 F_0 パターンにおける推定誤差に関する客観評価（**RMSE**）

提案手法における潜在変数を藤崎モデルのフレーズおよびアクセント成分を表現する変数と仮定し学習したのち，エンコーダへ F_0 パターンを入力し各成分を推定および推定された各成分をデコーダへ入力し F_0 パターンを再構成した．表 2 より，提案手法が従来法よりも各特徴量を高精度に再現できていることがわかる．

4.4　逆問題の解法にかかる処理時間

提案手法における潜在変数を藤崎モデルのフレーズおよびアクセント成分を表現する変数と仮定し学習したのち，エンコーダへ F_0 パターンを入力し各成分を推定した．表 3 より，モデル構造に CNN を用いた提案手法が従来法よりも高速に推定できることがわかる．

5　おわりに

本稿では，F_0 パターン生成過程のための深層生成モデルを提案した．実験的評価結果より，音声および歌声の F_0 パターンに内在するパラメータから F_0 パターンを推定，および，その逆問題を高精度かつ高速に解くことを可能であることを示した．今後は，提案手法に基づき，F_0 パターンの変換や転写を行う予定である．

謝辞　本研究の一部は，JSPS 科研費 17H01763 の助成を受け実施したものである．

参考文献

[1] Hiroya Fujisaki, *Vocal physiology: Voice production, mechanisms and functions*, pp. 347–355, 1988.
[2] Hirokazu Kameoka *et al.*, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1042–1053, 2015.
[3] Siu Wa Lee *et al.*, in *ICASSP*, 2012, pp. 429–432.
[4] Yasunori Ohishi *et al.*, in *INTERSPEECH*, 2012, pp. 474–477.
[5] Diederik P Kingma *et al.*, *arXiv preprint arXiv:1312.6114*, 2013.
[6] Casper Kaae Sønderby *et al.*, in *NIPS*, 2016, pp. 3738–3746.