DNN-SPACE：テキスト情報を利用した音声 F_0 パターン生成過程の確率モデル*

◎北条伸克 (NTT)，大杉康仁 (東大)，井島勇祐，亀岡弘和 (NTT)

1 はじめに

音声の基本周波数パターン（F_0 パターン）には，統語構造等の言語的な情報や，話者性，感情，意図などの非言語的な情報が豊富に含まれることが知られている．したがって，音声アプリケーションで F_0 を扱う場合，F_0 パターンのモデル化は非常に有用である．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り決定されるため，この物理的な制約を模した生成モデルが望まれる．藤崎モデル [1] は，F_0 の生成過程のモデルであり，生理学的・言語学的に意味のある少数のパラメータを用いて実測の F_0 パターンを良く近似できることが知られている．従来研究 [2] では，藤崎モデルをベースにした F_0 パターン生成過程の確率モデル（Statistical Phrase/Accent Command Estimation; SPACE）が提案されており，観測 F_0 パターンから藤崎モデルの指令列を推定するためのアルゴリズムが導出されている．

藤崎モデルの指令列は，発話の言語学的情報とも密接な関係にあるパラメータである．したがって，観測 F_0 パターンに加え，発話の言語学的情報をも利用し，藤崎モデル指令列推定を行うことができれば，推定精度の向上が可能であると考えられる．本研究では，言語学的情報を活用し，藤崎モデル指令列推定精度を向上させることを目的とし，言語学的情報と藤崎モデル指令列の関係を DNN によりモデル化し，SPACE と統合する．動作実験により，提案モデルにより，従来法の指令列推定誤りが修正される効果を示す．

2 音声 F_0 パターンの確率モデル

2.1 藤崎モデルの確率モデル化（SPACE）

本章では，まず従来研究の音声 F_0 パターンの確率モデル [2] について概説する．藤崎モデルは，対数 F_0 パターン $y[k]$ を，フレーズ成分 $x_p[k]$，アクセント成分 $x_a[k]$，ベースライン成分 x_b の和によりモデル化する．フレーズ成分とアクセント成分は，それぞれフレーズ指令と呼ばれるパルス波の列 $u_p[k]$ とアクセント指令と呼ばれる矩形波の列 $u_a[k]$ を入力とした臨界制動の二次線形系により表現される．

$$y[k] = x_p[k] + x_a[k] + x_b \tag{1}$$

$$x_p[k] = G_p[k] * u_p[k] \tag{2}$$

$$x_a[k] = G_a[k] * u_a[k] \tag{3}$$

ここで $*$ は畳み込みの演算子，$G_p[k], G_a[k]$ をそれぞれの二次線形系の応答関数とした．

音声 F_0 パターンの確率モデル [2] では，各指令列を，下記のパラメータセットを持つ経路制約付き HMM の出力系列としてモデル化することにより，各指令関数の形状の制約や，両指令が同時に生起されないという制約を表現することができる．

出力系列：$\boldsymbol{o}[k] = (u_p[k], u_a[k])^{\mathrm{T}}$

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：$\boldsymbol{s} = \{s_k\}_k$

出力分布：$p(\boldsymbol{o}[k]|s_k = i) = \mathcal{N}(\boldsymbol{o}[k]; \boldsymbol{c}_i, \Sigma_i)$

$$\boldsymbol{c}_i = \begin{cases} (0,0)^{\mathrm{T}} & (i \in r_0, r_1) \\ (u_p^{(m)}, 0)^{\mathrm{T}} & (i \in p_m) \\ (0, u_a^{(n)})^{\mathrm{T}} & (i \in a_n) \end{cases}, \Sigma_i = \begin{bmatrix} v_{p,i}^2 & 0 \\ 0 & v_{a,i}^2 \end{bmatrix}.$$

遷移確率：$\phi_{\hat{i},i} = p(s_k = i | s_{k-1} = \hat{i})$

さらに，指令列関数 $\boldsymbol{o}$ が与えられたもとでの観測 F_0 パターン $\boldsymbol{y}$ の生成確率を，

$$p(\boldsymbol{y}|\boldsymbol{o}) = \prod_{k=1}^{K} \mathcal{N}(y[k]; x[k], v_n^2[k]) \tag{4}$$

$$x[k] = G_p[k] * u_p[k] + G_a[k] * u_a[k] + x_b \tag{5}$$

によりモデル化する．

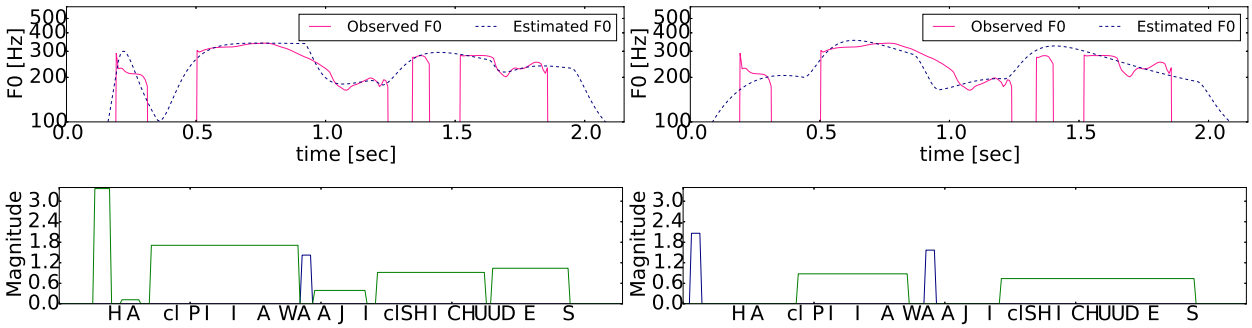
上述の藤崎モデルのパラメータ生成モデルの結合分布は，$p(\boldsymbol{y}, \boldsymbol{o}, \boldsymbol{s}, \theta) = p(\boldsymbol{y}|\boldsymbol{o})p(\boldsymbol{o}|\boldsymbol{s}, \theta)p(\boldsymbol{s})p(\theta)$ と書ける．ここで，上述の HMM の状態出力のパラメータを θ とした．従来研究 [2] では，補助関数法および EM アルゴリズムに基づき，与えられた $\boldsymbol{y}$ に対し，$\boldsymbol{s}$ を周辺化し，$p(\boldsymbol{o}, \theta|\boldsymbol{y})$ を最大化する $\boldsymbol{o}, \theta$ を求めるアルゴリズムが提案されている．また，従来研究 [3] では，パラメータ推定の精度向上と高速化のため，EM アルゴリズムの E ステップ (Forward–Backward アルゴリズムによる状態経路の事後確率計算) を Viterbi アルゴリズムによる状態経路最尤推定に置き換え，$p(\boldsymbol{o}, \boldsymbol{s}, \theta|\boldsymbol{y})$ を最大化するアルゴリズムが提案され，その効果が示されている．

2.2 DNN-SPACE

提案法は，言語特徴量系列を $\boldsymbol{w} = \{\boldsymbol{w}[k]\}_k$ としたとき，HMM 状態系列 $\boldsymbol{s}$ から $\boldsymbol{w}$ が生成される過程 $p(\boldsymbol{w}|\boldsymbol{s}, \lambda)$ をモデル化し，SPACE と統合する．ただし，λ は後述の DNN のパラメータとする．$p(\boldsymbol{w}|\boldsymbol{s}, \lambda)$ を利用すると，観測 F_0 パターン及び言語特徴量の生成過程は，結合分布 $p(\boldsymbol{y}, \boldsymbol{o}, \boldsymbol{w}, \boldsymbol{s}, \theta|\lambda) = p(\boldsymbol{y}|\boldsymbol{o})p(\boldsymbol{o}|\boldsymbol{s}, \theta)p(\boldsymbol{w}|\boldsymbol{s}, \lambda)p(\boldsymbol{s})p(\theta)$ によりモデル化される．パラメータ推定時には，$\boldsymbol{y}, \boldsymbol{w}, \lambda$ が与えられたとき，$p(\boldsymbol{o}, \boldsymbol{s}, \theta|\boldsymbol{y}, \boldsymbol{w}, \lambda)$ を最大化する $\boldsymbol{o}, \boldsymbol{s}, \theta$ を得る．これにより，言語特徴量系列 $\boldsymbol{w}$ との関係を考慮した上で，指令列の推定を行う．

テキスト音声合成の分野では，正確な F_0 パターンの予測のためには，例えば決定木やニューラルネットワーク等により，音素やアクセント情報などの多様な言語学的情報の依存関係をモデル化することが有用であることが知られている．したがって，提案法においても同様に，多様な言語学的情報を変数 $\boldsymbol{w}$ で表現し，$p(\boldsymbol{w}|\boldsymbol{s}, \lambda)$ により柔軟にモデル化することが有用であると考えられる．ここで，$\boldsymbol{w}$ を言語学的情報を表現する数値ベクトルとした場合，$p(\boldsymbol{w}|\boldsymbol{s}, \lambda)$ として，

*DNN-SPACE : The generative model of speech F_0 contour using linguistic information. by HOJO, Nobukatsu (NTT), OHSUGI, Yasuhito (The University of Tokyo), IJIMA,Yusuke, KAMEOKA, Hirokazu (NTT)



(1) 従来法　　(2) 提案法

Fig. 1　パラメータ推定結果の例．上段実線は観測 F_0 パターン，点線は推定 F_0 パターン．下段は推定されたフレーズ指令（青），アクセント指令（緑）．

GMM のように，次元間の依存関係が比較的単純な分布を仮定すると，多様な言語学的情報の依存関係を十分に表現できない可能性がある．以上の観点から，提案法では，

$$p(\boldsymbol{w}|\boldsymbol{s},\lambda) = \prod_k p(\boldsymbol{w}[k]|s_k,\lambda) \tag{6}$$

$$= \prod_k \frac{p(s_k|\boldsymbol{w}[k],\lambda)p(\boldsymbol{w}[k])}{p(s_k)} \tag{7}$$

のように変形し，$\boldsymbol{w}[k]$ を入力，HMM 状態 s_k の事後確率を出力とする DNN により $p(s_k|\boldsymbol{w}[k],\lambda)$ をモデル化する．$p(s_k)$ には，各状態の相対頻度を使用し，$p(\boldsymbol{w}[k])$ は定数と仮定する．提案法では，$p(s_k|\boldsymbol{w}[k],\lambda)$ において，$\boldsymbol{w}$ を DNN の入力ベクトルとして使用することで，$\boldsymbol{w}$ の次元間の依存関係を柔軟にモデル化し，F_0 パターンの生成モデルに組み込むことが可能である．

提案モデルのパラメータ推定アルゴリズムは，$p(\boldsymbol{o},\boldsymbol{s},\theta,\lambda|\boldsymbol{y},\boldsymbol{w})$ を最大化する．ここで，従来研究 [3] と同様に，$\boldsymbol{o}$ を固定した上で $\boldsymbol{s}$ を更新するステップと，$\boldsymbol{s}$ を固定した上で $\boldsymbol{o}$ を更新するステップを反復する．この際，$\boldsymbol{o}$ を更新するステップについては，従来研究 [3] と同様である．また，$\log p(\boldsymbol{y},\boldsymbol{o},\boldsymbol{w},\boldsymbol{s},\theta|\lambda)$ のうち，$\boldsymbol{s}$ に依存する項は $\log p(\boldsymbol{o}|\boldsymbol{s},\theta)+\log p(\boldsymbol{w}|\boldsymbol{s},\lambda)+\log p(\boldsymbol{s})$ であるが，式 (6) により，$\log p(\boldsymbol{w}|\boldsymbol{s},\lambda)$ の時刻独立性を仮定したことから，$\log p(\boldsymbol{o}|\boldsymbol{s},\theta), \log p(\boldsymbol{w}|\boldsymbol{s},\lambda)$ はそれぞれ時刻 k ごとに分解可能である．したがって，$\boldsymbol{s}$ の更新ステップは，従来研究 [3] と同様に，Viterbi アルゴリズムに基づき導出可能である．

3　実験

3.1　実験条件

本章では，提案法により，言語学的情報を利用することで指令列推定精度が向上するという仮定の妥当性を検証するために行った動作実験について述べる．本実験では，音声データベースとして，女性 1 名のプロナレータによる，読み上げ調の発話を使用した．言語特徴量 $\boldsymbol{w}$ として，DNN 音声合成 [4] 等で使用される言語特徴量ベクトルと同様の，音素情報，当該アクセント句のモーラ数，当該アクセント句のアクセント型などからなる 495 次元のベクトルを使用した．DNN のパラメータ θ の学習のため，4358 文の音声に対し，従来法 [2] により藤崎モデル指令列の推定を行い，推定結果を正解データとした．この際，正解データ中の各指令を，その大きさによりクラスタリングし，それぞれ 20 状態に分割した．指令なしの状態は，フレーズ指令を持つ状態，アクセント指令を持つ状態の場合を区別した．以上から，フレーズ指令 20 状態，アクセント指令 20 状態，指令なし 2 状態の全 42 状態の識別器として，DNN の学習を行った．続いて，DNN の学習データに含まれない音声に対し，学習された DNN を使用した提案法により，言語特徴量と観測 F_0 パターンから，藤崎モデル指令列の推定を行った．DNN の隠れ層数は 2 層，ユニット数が 512 であるフルコネクトのネットワークを使用し，活性化関数はシグモイド関数とした．

3.2　実験結果

提案法，従来法によるパラメータ推定結果の例（文章：「ハッピーアワー実施中です（ハッピーアワー ジッシチュウデス）」）を図 1 に示した．時刻 0.1 [sec] 付近の指令列は，フレーズ先頭付近であるため，フレーズ指令が妥当であると考えられるものの，従来法では，アクセント指令が推定されており，誤推定が生じていると考えられる．一方，提案法では，DNN の事後確率から，フレーズ先頭付近である，という言語学的情報に基づき，フレーズ指令が推定されている．以上より，言語学的情報のモデルを統合することにより，推定誤りが減少する例が確認された．

4　むすび

本稿では，観測 F_0 パターンから藤崎モデル指令列を推定する従来法 [2] に，さらに言語学的情報のモデルを統合することで，観測 F_0 パターンと，対応する言語学的情報から藤崎モデル指令列を推定する手法を提案した．動作実験により，言語特徴量を使用することで，フレーズ指令とアクセント指令の誤推定などのパラメータ推定誤りが減少する例を確認した．今後は，専門家によるフレーズ指令・アクセント指令の正解ラベル付き音声データに提案法を適用し，パラメータ推定精度を客観的に評価する予定である．

参考文献

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

[2] H. Kameoka et al., “Generative modeling of voice fundamental frequency contours,” IEEE/ACM Trans. Audio, Speech, and Language, Vol. 23, No. 6, pp. 1042–1053, Jun. 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,” 情報処理学会音声言語情報処理研究会, pp.1–6, Nov. 2016.

[4] H.Zen et al., “Statistical parametric speech synthesis using deep neural networks, ” Proc. ICASSP, pp.7962–7966, 2013.