T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	観測値の不連続性を考慮したHMMに基づくF0モデル化の検討
Title(English)	A study on F0 modeling based on discontinuous observation HMM
著者(和文)	郡山 知樹, 能勢 隆, 小林隆夫
Authors(English)	Tomoki Koriyama, Takashi Nose, Takao Kobayashi
出典(和文)	日本音響学会2012年春季研究発表会講演論文集, Vol. , No. , pp. 305-306
Citation(English)	Proceedings of the ASJ 2012 Spring Meeting, Vol. , No. , pp. 305-306
発行日 / Pub. date	2012, 3

Powered by T2R2 (Science Tokyo Research Repository)

観測値の不連続性を考慮したHMMに基づくF0モデル化の検討*

☆郡山知樹，能勢　隆，小林隆夫 (東工大)

1　はじめに

多空間確率分布に基づくHMM(MSD-HMM)[1]は，有声・無声区間を含む基本周波数(F0)パタンをモデル化可能な手法として，HMM音声合成において広く使用されている．しかしHMMの基本単位として，音素ではなくアクセント句や文献[2]で提案した韻律ラベルに基づく単位を使用して，アクセントやイントネーションを表すF0の概形のモデル化を目的とした場合，空間重みをF0と同時にモデル化するMSD-HMMは必ずしも最適とは言えない．そこで本研究では，有声の単一空間確率分布から成るF0モデルにより，F0の概形をモデル化する手法を提案する．本稿では提案法のモデルパラメータの推定方法を示し，単語音声を用いた実験によりMSD-HMMによるモデル化との比較を示す．

2　F0概形モデルのためのHMM

MSD-HMMは次元の異なる複数の特徴が混合した観測系列を，それぞれの空間における確率分布によってモデル化するHMMであり，F0パタンに対しては有声区間を連続値，無声区間を単一の離散シンボルとした確率分布が用いられる．F0モデルに対するMSD-HMMは$\lambda = \{\boldsymbol{A}, \boldsymbol{w}, \boldsymbol{M}, \boldsymbol{V}\}$で表される．ただし$\boldsymbol{A}$は状態遷移確率，$\boldsymbol{w}$は空間重み，$\boldsymbol{M}$，$\boldsymbol{V}$はそれぞれ有声空間の平均と共分散行列を表す．HMM音声合成では，音素単位のMSD-HMMを用いることで，音素の各状態におけるF0値と有声・無声らしさを同時にモデル化することができる．

これに対し，我々はアクセントによるF0の下降区間などの韻律ラベルに基づく韻律単位をHMMの基本単位として用いることで，合成音声の自然性を維持したままF0モデルのパラメータ数を大幅に削減可能であることを報告した[2]．この手法は有声・無声に関わらず連続的な「F0概形」が韻律単位毎に決まり，F0値が観測されるか否かは音素などのコンテキストに起因するものであるという仮定に基づいている．F0概形は有声の単一空間のみで表されることから，本研究では有声の単一空間ガウス分布で構成されるHMM，$\lambda = \{\boldsymbol{A}, \boldsymbol{M}, \boldsymbol{V}\}$のモデル化手法を提案する．なお，この手法は韻律単位だけでなくアクセント句のF0概形のモデル化など任意の区間でのモデル化を目的としている．

F0パタンは無声区間では値が観測されず局所的に不連続となるため，F0概形を通常のHMMで直接モデル化することはできない．過去の研究では，F0抽出器の第一候補や何らかの補間手法を用いて無声区間を補間し，得られた系列に対する尤度を最大化することによって学習する手法が提案されている[3, 4]が，この場合モデル化性能が補間手法の性能に依存してしまう．これに対し提案法では，有声区間の観測系列に対して尤度を最大化することによってモデルを学習する．このとき，文献[5]で提案された無声区間を隠れ変数として考慮する手法を用いる．これによって，観測されたF0を重視したF0概形のモデル化が可能になる．

2.1　学習アルゴリズム

F0系列を$\boldsymbol{O} = [\boldsymbol{o}_1 \dots \boldsymbol{o}_T]$とする．ただし$\boldsymbol{O}$は単一空間から成り，無声区間では値が隠れているものとする．F0の観測された有声フレームの集合を$S^{(v)}$，観測されなかった無声フレームの集合を$S^{(u)}$とし，有声区間，無声区間のF0系列をそれぞれ$\boldsymbol{O}^{(v)}, \boldsymbol{O}^{(u)}$で表す．ここで，実際の有声観測系列である$\boldsymbol{O}^{(v)}$の尤度

$$P(\boldsymbol{O}^{(v)}|\lambda) = \int P(\boldsymbol{O}^{(u)}, \boldsymbol{O}^{(v)}|\lambda) d\boldsymbol{O}^{(u)} \tag{1}$$

を最大化することを考える．パラメータ推定には通常のHMMと同様にEMアルゴリズムを用いることが可能であり，EMアルゴリズムにおいて隠れ変数に無声区間のF0系列$\boldsymbol{O}^{(u)}$を使用することにより，Q関数は以下のように表される．

$$Q(\lambda, \tilde{\lambda}) = \mathbb{E}_{\boldsymbol{q}, \boldsymbol{O}^{(u)}} \left[\log P(\boldsymbol{q}, \boldsymbol{O}^{(u)}, \boldsymbol{O}^{(v)}|\tilde{\lambda}) \right] \tag{2}$$

これを状態遷移確率a_{ij}，出力確率$b_i(\boldsymbol{o})$に対するQ関数$Q(\lambda, \tilde{a})$，$Q(\lambda, \tilde{b})$に分解すると，

$$Q(\lambda, \tilde{a}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T-1} \xi_t^{(v)}(i,j) \log \tilde{a}_{ij} \tag{3}$$

$$Q(\lambda, \tilde{b}) = \sum_{i=1}^{N} \left\{ \sum_{t \in S^{(u)}} \int b_i(\boldsymbol{x}) \gamma_t^{(v)}(i) \log \tilde{b}_i(\boldsymbol{x}) d\boldsymbol{x} + \sum_{t \in S^{(v)}} \gamma_t^{(v)}(i) \log \tilde{b}_i(\boldsymbol{o}_t) \right\} \tag{4}$$

となる．ただし，Nは状態数，Tは総フレーム数，$\boldsymbol{q} = [q_1, \dots, q_T]$は状態系列である．$\gamma_t^{(v)}(i)$は状態占有確率であり，$\gamma_t^{(v)}(i)$，$\xi_t^{(v)}(i,j)$はそれぞれ

$$\begin{aligned} \gamma_t^{(v)}(i) &= P(q_t = i|\boldsymbol{O}^{(v)}, \lambda) \\ &= \frac{\alpha_t^{(v)}(i) \beta_t^{(v)}(i)}{\sum_{k=1}^{N} \alpha_t^{(v)}(k) \beta_t^{(v)}(k)} \end{aligned} \tag{5}$$

$$\begin{aligned} \xi_t^{(v)}(i,j) &= P(q_t = i, q_{t+1} = j|\boldsymbol{O}^{(v)}, \lambda) \\ &= \frac{\alpha_t^{(v)}(i) a_{ij} b'_j(\boldsymbol{o}_{t+1}) \beta_{t+1}^{(v)}(j)}{P(\boldsymbol{O}^{(v)}|\lambda)} \end{aligned} \tag{6}$$

で表される．ただし，

$$b'_i(\boldsymbol{o}_t) = \begin{cases} b_i(\boldsymbol{o}_t) & t \in S^{(v)} \\ 1 & t \in S^{(u)} \end{cases} \tag{7}$$

であり，$\alpha_t^{(v)}(i)$，$\beta_t^{(v)}(i)$はフォーワード・バックワードアルゴリズムによって以下のように求められる．

$$\alpha_1^{(v)}(i) = \pi_i b'_i(\boldsymbol{o}_T) \tag{8}$$

$$\beta_T^{(v)}(i) = 1 \tag{9}$$

---

*A study on F0 modeling based on discontinuous observation HMM, by KORIYAMA, Tomoki, NOSE, Takashi, and KOBAYASHI, Takao (Tokyo Institute of Technology)

$$\alpha_t^{(v)}(i) = \left(\sum_{j=1}^{N} \alpha_{t-1}^{(v)}(j) a_{ji} \right) b_i'(\boldsymbol{o}_t) \tag{10}$$

$$\beta_t^{(v)}(i) = \sum_{j=1}^{N} a_{ij} b_j'(\boldsymbol{o}_{t+1}) \beta_{t+1}^{(v)}(j) \tag{11}$$

Q 関数を最大にすることでモデルパラメータの更新を行う．ここでは簡単のため，出力確率 $b_i(\boldsymbol{o})$ が単一正規分布 $\mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_i, \boldsymbol{V}_i)$ であるとする．モデルパラメータの更新式は以下の通りに表される．

$$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t^{(v)}(i,j)}{\sum_{k=1}^{N} \sum_{t=1}^{T-1} \xi_t^{(v)}(i,k)} \tag{12}$$

$$\tilde{\boldsymbol{\mu}}_i = \frac{\sum_{t \in S^{(v)}} \gamma_t^{(v)}(i) \boldsymbol{o}_t + \sum_{t \in S^{(u)}} \gamma_t^{(v)}(i) \boldsymbol{\mu}_i}{\sum_{t=1}^{T} \gamma_t^{(v)}(i)} \tag{13}$$

$$\tilde{\boldsymbol{V}}_i = \frac{1}{\sum_{t=1}^{T} \gamma_t^{(v)}(i)} \left\{ \sum_{t \in S^{(v)}} \gamma_t^{(v)}(i) (\boldsymbol{o}_t - \tilde{\boldsymbol{\mu}}_i)(\boldsymbol{o}_t - \tilde{\boldsymbol{\mu}}_i)^\top + \sum_{t \in S^{(u)}} \gamma_t^{(v)}(i) \left(\boldsymbol{V}_i + (\boldsymbol{\mu}_i - \tilde{\boldsymbol{\mu}}_i)(\boldsymbol{\mu}_i - \tilde{\boldsymbol{\mu}}_i)^\top \right) \right\} \tag{14}$$

3 実験

提案手法の特徴を調べるため，単語音声を用いた実験を行った．使用したデータは ATR 日本語音声データベース C-set に含まれる話者 f104 のアクセント型が中高型である 20 単語である．特徴ベクトルは STRAIGHT で抽出した対数 F0 の 1 次元ベクトルとし，モデルは 3 状態の left-to-right HMM とした．

表 1，2 に MSD-HMM および提案法のモデルパラメータを学習した結果を示す．表中の $w_i^{(v)}$ は有声空間の重みである．また，図 1 に「いよいよ」「ぽけっと」の学習データの F0 パタンと状態占有確率の時間変化を示す．図の「いよいよ」の状態占有確率の結果から，HMM の第 1 状態はピッチ上昇部分，第 2 状態はピッチ停滞部分，第 3 状態はピッチ下降部分になっていることがわかる．「いよいよ」の場合は無声区間がないため提案法の結果は MSD-HMM とほぼ等しくなっている．一方「ぽけっと」は「け」の前と「と」の前に無声区間があり，状態占有確率が異なる結果となっている．MSD-HMM では，前者の無声区間は第 1 状態に，後者の無声区間は第 3 状態にほぼ完全に占有されている．このようになった理由は第 2 状態に比べ，第 1 状態，第 3 状態の無声空間重みが大きいからである．提案法では無声区間がどちらか一方の状態に占有されることはなくなり、この違いが，状態遷移確率の相違の原因となっている．このことから提案法のモデル学習法は，MSD-HMM に比べ無声区間の影響を受けにくいものと考えられる．

4 まとめ

本稿ではアクセントやイントネーションを表す F0 概形をモデル化する手法として，無声区間を隠れ変数として使用し有声区間の観測系列の尤度を最大化することによるモデル学習法を提案した．モデルパラメータの推定アルゴリズムを示し，さらにアクセント句に対してモデル化実験を行い，提案法が F0 概形のモデル化手法として有効である可能性を示した．今後の課題としては実際の音声合成に適用することによって，モデルの有効性を検討することが挙げられる．

Table 1　MSD-HMM のモデルパラメータ

i	1	2	3
$w_i^{(v)}$	0.706	0.957	0.822
μ_i	5.34	5.54	5.13
σ_i	0.0700	0.0560	0.188
$a_{i,i+1}$	0.0249	0.0144	0.0176

Table 2　提案 HMM のモデルパラメータ

i	1	2	3
μ_i	5.34	5.54	5.14
σ_i	0.0706	0.0542	0.191
$a_{i,i+1}$	0.0277	0.0136	0.0175

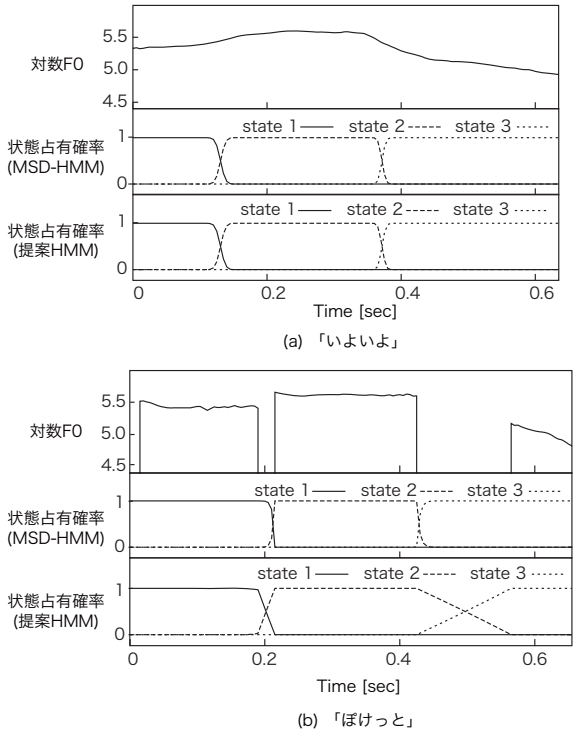

Fig. 1　中高型単語における状態占有確率の時間変化

謝辞　本研究の一部は，日本学術振興会科学研究費補助金 (課題番号 21300063，23700195) の助成を得た．

参考文献

[1] 徳田他，“多空間上の確率分布に基づいた HMM，” 信学論 (D-II)，vol.J83-D-II，no.7，pp.1579–1589, 2000.

[2] 郡山他，“韻律イベント HMM を用いた対話音声 F0 生成，” 信学技報，vol.111，no.365，SP2011-98，pp.185–190，2011.

[3] Q. Zhang, et al., “Improved modeling for F0 generation and V/U decision in HMM-based TTS,” Proc. ICASSP 2010, pp.4606–4609, 2010.

[4] K. Yu, et al., “Continuous F0 Modeling for HMM Based Statistical Parametric Speech Synthesis,” IEEE Trans. Audio, Speech & Lang., vol.19, no.5, pp.1071–1079, 2011.

[5] 亀岡他，“音声 F0 パターン生成過程の確率モデル，” 音講論（秋），1-1-3，pp.207–210，2010.