確率生成モデルに基づく連続音声からの教師なし音素・単語・文法獲得

落合翔馬 [1]　長野匡隼 [1]　中村友昭 [1]
[1] 電気通信大学
o1910151@edu.cc.uec.ac.jp

概要

人間は，二重分節構造を持つ連続音声信号を教師なしで音素や単語に分割し，文法を学習することができる．ロボットが人間のように言語を獲得するには，そのような二重分節構造の学習が可能なモデルが必要となる．本稿では連続音声から音素・単語・文法を学習可能な確率的生成モデルを提案する．このモデルは，音素を学習する Gaussian Process Hidden Semi Markov Model (GP-HSMM) と，単語・文法を学習する Hidden Semi Markov Model (HSMM) で構成された二階層のモデルである．実験では，提案手法によって連続音声から音素・単語・文法を学習できることを示す．

1　はじめに

人間の幼児は正解を与えられなくとも，教師なしで連続音声信号から言語を学習している．言語は二重の分節構造を持っており，連続音声を分節化することで音素を，音素を分節化することで単語・文法を学習することができる．そのような人間の言語学習能力を持つロボットを実現するためには，二重分節構造を持つ時系列データを教師なしで分節化可能なモデルが必要である．音声認識の分野では，大量の音声データやラベル付けされたコーパスを用いた教師あり学習によって，音声を分割して認識する手法が主に用いられている [1][2][3][4]．しかし，人間は音声信号から音素・単語・文法を学習するために大量のラベル付きデータセットやコーパスを用いておらず，このような手法は人間の言語学習とは異なっている．一方，教師なしで連続音声から音素と単語を学習するモデルとして，Nonparametric Bayesian Double Articulation Analyzer (NPB-DAA）が提案されている [5]．また，我々は Gaussian Process Hidden Semi Markov Model (GP-HSMM)[6][7] と Hidden Semi Markov Model (HSMM) を組み合わせることで，二重の分節構造を持つ連続音声から音素と単語の学習が可能な確率的生成モデル GP-HSMM-based Double Articulation Analyzer（GP-HSMM-DAA）を提案した [8]．しかしこれらの手法では，音素と単語の学習に留まっており文法の学習までは実現できていない．

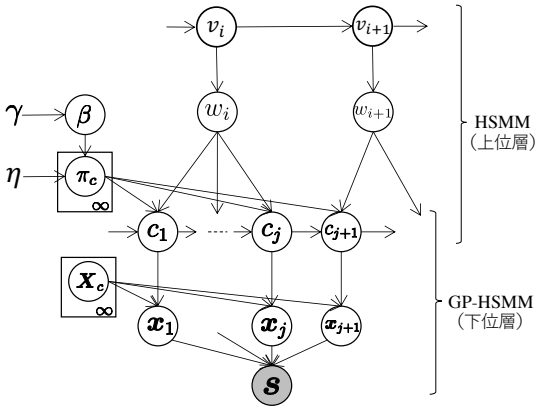

図 1　提案手法のグラフィカルモデル

そこで本稿では二重分節構造を持つ連続音声から音素と単語だけでなく，文法も教師なしで学習可能な確率的生成モデルを提案する．提案するモデルは GP-HSMM-DAA と同様に GP-HSMM と HSMM から構成されている階層的なモデルである．この階層構造により，教師なしで音声信号の二重分節構造を学習することが可能である．実験では日本語の音声データから，教師なしで音素，単語そして文法を学習可能であることを示す．

2　提案手法

図 1 が提案手法のグラフィカルモデルであり，下位の層が GP-HSMM，上位の層が HSMM で構成された確率的生成モデルである．GP-HSMM を用いて音声信号から音素列を学習し，HSMM を用いて音素

列から単語と文法を学習する．

2.1 生成過程

品詞に相当する単語クラス v_i は直前のクラス v_{i-1} によって生成される．

$$v_i \sim P(v|v_{i-1}) \tag{1}$$

この単語クラスの遷移規則が文法であり，従来の GP-HSMM-DAA[8] とは異なる部分である．次に，単語クラス v_i に従い単語 w_i が生成される．

$$w_i \sim P(w|v_i) \tag{2}$$

また，単語 w_i を構成している音素クラス c_j は直前の音素クラス c_{j-1} と遷移確率 π_c によって生成される．

$$c_j \sim P(c|c_{j-1}, \pi_{c_{j-1}}, w_i) \tag{3}$$

遷移確率 π_c の生成には階層ディリクレ過程 (HDP) を用い，以下のように Stick-breaking Process によって生成された β を基底測度とした Dirichlet process(DP) によって生成される．

$$\beta \sim \mathrm{GEM}(\gamma) \tag{4}$$

$$\pi_c \sim \mathrm{DP}(\eta, \beta) \tag{5}$$

これによりデータの複雑さに応じて，自動的に音素クラスの数を推定することが可能となる．音素クラス c の音声信号 $\boldsymbol{x}_j$ はガウス過程から生成される．

$$\boldsymbol{x}_j \sim \mathcal{GP}(\boldsymbol{x}|\boldsymbol{X}_c) \tag{6}$$

ただし，$\boldsymbol{X}_c$ は音素クラス c のガウス過程のパラメータである．このように生成された各音素の音声信号を連結することで連続音声信号 $\boldsymbol{S} = \{\boldsymbol{x}_1, \cdots \boldsymbol{x}_J\}$ は生成される．

2.2 ガウス過程

提案手法の下位層では，単位系列 $\boldsymbol{x}$ 内のタイムステップ t における出力 x_t を連続的な軌道として表現するためにガウス過程回帰を用いる．ガウス過程回帰では同じクラスに属するタイムステップ t における出力 x_t の複数のペア $(\boldsymbol{t}, \boldsymbol{X})$ が得られた時，タイムステップ $\hat{t}$ における出力 $\hat{x}$ の予測分布は以下のガウス分布となる．

$$p(\hat{x}|\hat{t}, \boldsymbol{X}, \boldsymbol{t}) \propto \mathcal{N}(\boldsymbol{k}^T\boldsymbol{C}^{-1}\boldsymbol{X}, k(\hat{t},\hat{t}) - \boldsymbol{k}^T\boldsymbol{C}^{-1}\boldsymbol{k}) \tag{7}$$

ただし，$k(\cdot,\cdot)$ はカーネル関数である．C は，$\boldsymbol{t}$ の p 番目と q 番目の要素を t_p, t_q とした時，p 行 q 列の値が

$$C(t_p, t_q) = k(t_p, t_q) + \phi^{-1}\delta_{pq}, \tag{8}$$

Algorithm 1 相互更新による学習

```
// Initialization
Set P(C|W) to uniform distribution

for m = 1 to M do
    // Learning of lower layer
    C ~ GP − HSMM(S, P(C|W))

    // Learning of higher layer
    V, W ~ HSMM(C)

    // Parameter update
    Update P(C|W) from W
end for
```

となる行列である．ϕ は観測値に含まれるノイズの表すハイパーパラメータである．また，$\boldsymbol{k}$ は $k(t_p, \hat{t})$ を p 番目の要素に持つベクトルである．本稿ではカーネル関数として以下の式を用いる．

$$k(t_p, t_q) = \theta_0 \exp(-\frac{1}{2}\theta_1||t_p - t_q||^2) + \theta_2 + \theta_3 t_p t_q \tag{9}$$

θ_* はカーネルのハイパーパラメータである．出力値が多次元のベクトル $\boldsymbol{x}_t = (x_t^{(1)}, \cdots, x_t^{(d)}, \cdots)$ の場合は各次元が独立に生成されると仮定し，時刻 t の観測値 $\boldsymbol{x}$ がクラス c に対応するガウス過程から生成される確率 $\mathcal{GP}(\boldsymbol{x}|\boldsymbol{X}_c)$ を以下のように計算する．

$$\mathcal{GP}(\boldsymbol{x}|\boldsymbol{X}_c) = \prod_d^D p(x_t^{(d)}|t, \boldsymbol{X}_c^{(d)}) \tag{10}$$

2.3 パラメータの推論

提案モデルは二階層のモデルであり，単純にはパラメータを推論することが困難である．そこで，各階層を交互に推論することで，モデル全体のパラメータを最適化する．Algorithm 1 が相互学習を用いたパラメータ推定のアルゴリズムである．

まず下位層において，観測された音声信号 S を GP-HSMM により分節化し音素クラス系列 C をサンプリングする．次に，得られた音素クラス系列を，上位層の HSMM によって分節化することで，単語系列 W と単語クラス系列 V をサンプリングする．上位層では分節化された単語 w から音素クラス c が生成される条件付確率 $P(c|w)$ を計算し，下位層 (GP-HSMM) に送る．GP-HSMM では受け取った $P(c|w)$ を音素の事前分布として用い，再度音素クラスのサンプリングを行う．この相互更新を M 回繰り返すことによってパラメータの最適化をする．

GP-HSMM と HSMM では分節長とクラスを効率的にサンプリングするために Forward Filtering

- Backward Sampling アルゴリズムを用いる．GP-HSMM の Frward Filtering では，音声信号のタイムステップ t を終端とする長さ k の部分系列が音素クラス c となる前向き確率は次式のようになる．

$$\alpha_p[t][k][c] = \mathcal{G}P(\boldsymbol{x}_{t-k:t}|\boldsymbol{X}_c)P(c|w_i)P_{len}(k|\lambda_p) \times \sum_{k'=1}\sum_{c'=1} P(c|c',\pi_{c'})\alpha_p[t-k][k'][c'] \quad (11)$$

ただし，$P_{len}(k|\lambda_p)$ は分節長を決める λ_p をパラメータとするポアソン分布であり，遷移確率の計算には Product of Experts (PoE) 近似を用いて，$P(c|c',\pi_{c'},w_i) \approx P(c|c',\pi_{c'})P(c|w_i)$ とした．$P(c|w_i)$ は，上位層で計算された単語から音素クラス c が発生する確率であり，この確率により単語の持つ言語的な制約を，音素の学習に与えることができる．この前向き確率から音素クラス系列 C をサンプリングする．

次に上位層では，音素クラス系列 C を分節化することで，単語と単語クラスをサンプリングする．Foward Filtering では，音素のタイムステップ j を終端として，長さ k の部分系列が単語となり，そのクラスが v となる確率は次式のようになる．

$$\alpha_w[j][k][v] = P(C_{j-k:j}|v)P_{len}(k|\lambda_w) \times \sum_{k'=1}\sum_{v'=1} P(v|v')\alpha_w[j-k][k'][v'] \quad (12)$$

この前向き確率から単語系列 W と単語クラス系列 V をサンプリングすることができる．サンプリングされた W から $P(c|w_i)$ を更新し，GP-HSMM の計算に利用する．

以上のように，以下の手順を繰り返すことで，下位層と上位層が相互に影響しあい，音素・単語・文法を学習することができる．

1. 音声信号 S から音素クラス系列 C のサンプリング
2. 音素クラス系列 C から単語系列 W と単語クラス系列 V のサンプリング
3. 各単語から音素が発生する確率 $P(c|w_i)$ の更新

3 実験

提案手法の有効性を確認するために AIOI-dataset[1)] を改変したデータを用いて実験を行った．

1) https://github.com/EmergentSystemLabStudent/aioi_dataset

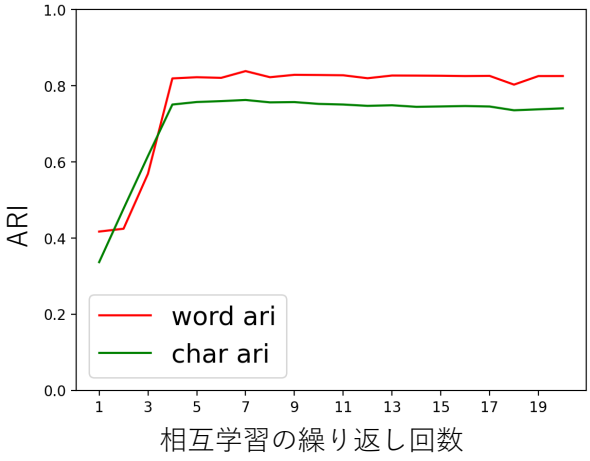

図 2 音素と単語の精度の推移

3.1 実験設定

AIOI-dataset は日本語の母音 {a,i,u,e,o} で構成された単語 {aioi,aue,ao,ie,uo} を組み合わせて作られた文を発話した音声データである．本実験では AIOI-dataset を単語毎に分割し，規則に従って 2 単語を組み合わせた 2 語文を新規に作成した．想定した文法規則は以下の通りである．

- 一単語目の出現単語: ao, uo, ie
- 二単語目の出現単語: aioi, aue, ie

すなわち，文は 9 種類 {“ao aioi”, “ao aue”, “ao ie”, “uo aioi”, “uo aue”, “uo ie”, “ie aioi”, “ie aue”, “ie ie”} である．これらを読み上げた音声データの数は 60 となった．観測系列として音声信号のメル周波数ケプストラム係数を Deep Sparse Auto encoder によって 3 次元に圧縮した特徴量を使用した．

単語のクラス数を 2，相互更新の繰り返し回数 $M = 20$ として，特徴量を分節化し文字・単語・文法の学習を行った．また，提案モデルは初期値の依存性があるため，学習を初期値を変えて 10 回試行し，単語の尤度が最も高い試行の結果を評価に用いた．評価指標には，正解と音素・単語の分類結果の Adjusted Rand Index (ARI) を用いた．ARI は，分類結果が正解に近いほど，1 に近い値となる指標である．

3.2 実験結果

図 2 の横軸が相互更新の繰り返し回数，縦軸が ARI である．また，赤線が単語の ARI であり，緑線が音素の ARI である．この図からパラメータの相互更新を行うごとに ARI が上昇しており，提案アルゴリズムが有効に働いていることが分かる．

表 1 各単語クラスに分類された単語．数字は音素クラスのインデックス，括弧内の文字は対応する音素である．

単語クラス 1	単語クラス 2
“87”（“ao”）	“802”（“aue”）
“07”（“uo”）	“8101”（“aioi”）
“12”（“ie”）	“12”（“ie”）

表 1 は，HSMM で各単語クラス ν に分類された単語（音素の部分系列）である．この結果と想定した文法規則を比べると，単語クラス 1 に一単語目に出現する単語，単語クラス 2 に二単語に出現する単語が分類されており，想定した文法規則が学習されている．

4 結論と今後の課題

本稿では，二重分節構造を持つ連続音声データから，教師なしで音素・単語・文法を学習するモデルを提案した．提案手法は，GP-HSMM と HSMM を組み合わせた二階層のモデルである．実験では AIOI-dataset から作成した文法の規則が存在するデータを用いて，音素・単語・文法が学習可能であることを示した．

現状の提案手法ではあらかじめ単語のクラス数を設定する必要があったため，今後は GP-HSMM と同様にクラス数を自動的に決定可能な，階層ディリクレ過程を導入する予定である．また，今回用いたデータは非常にシンプルな文法規則を想定した二語文であったため，今後はより複雑な文法規則の学習が可能か検証する予定である．

参考文献

[1] Tatsuya Kawahara, Akinobu Lee, Tetsunori Kobayashi, Kazuya Takeda, Nobuaki Minematsu, Shigeki Sagayama, Katsunobu Itou, Akinori Ito, Mikio Yamamoto, Atsushi Yamada, Takehito Utsuro, and Kiyohiro Shikano.“Free software toolkit for Japanese large vocabulary continuous speech recognition”, 6th International Conference on Spoken Language Processing, ICSLP 2000, 2000.

[2] George E. Dahl, Dong Yu, Li Deng, and Alex Acero. “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No.1 pp. 30-42, 2011.

[3] Hiroshi Seki, Kazumasa Yamamoto, and Seiichi Nakagawa. “Comparison of syllable-based and phoneme-based dnn-hmm in japanese speech recognition”. 2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA). pp. 249–254. 2014.

[4] Sei Ueno, Hirofumi Inaguma, Masato Mimura, and Tatsuya Kawahara. “Acoustic-to-word attention-based model complemented with character-level ctc-based model”. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5804–5808. 2018.

[5] Tadahiro Taniguchi, Ryo Nakashima, Hailong Liu, and Shogo Nagasaka.“Double articulation analyzer with deep sparse autoencoder for unsupervised word discovery from speech signals”.Advanced Robotics. Vol. 30, No.11-12 , pp. 770-783. 2016.

[6] Tomoaki Nakamura, Takayuki Nagai, Daichi Mochihashi, Ichiro Kobayashi, Hideki Asoh, and Masahide Kaneko. “Segmenting continuous motions with hidden semi-markov models and gaussian processes”, Frontiers in neurorobotics, Vol. 11, p. 67, 2017.

[7] Masatoshi Nagano, Tomoaki Nakamura, Takayuki Nagai, Daichi Mochihashi, Ichiro Kobayashi, and Masahide Kaneko. “ Sequence pattern extraction by segmenting time series data using gp-hsmm with hierarchical dirichlet process”, 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), p. 4067-4074, 2018.

[8] 長野匡隼，中村 友昭，“GP-HSMM に基づく二重分節化モデルによる連続音声の教師なし構造学習”，日本ロボット学会誌，2023