二重分節構造モデルを用いた連続音声からの教師なし音素・単語・文法獲得

落合翔馬 [1]　齋藤一誠 [1]　長野匡隼 [1]　中村友昭 [1]
[1] 電気通信大学
s_ochiai@radish.ee.uec.ac.jp

概要

人間は二重分節構造を持つ連続音声信号を明確な境界点やラベルなしに音素や単語に分割し，単語の遷移規則を文法として学習することが可能である．音声信号の二重分節構造を学習するモデルを構築することは，人間の言語獲得過程を構成論的に解明するために重要である．そこで本稿では，Gaussian Process Hidden Semi Markov Model (GP-HSMM) と Hidden Semi Markov Model (HSMM) を階層的に接続し，連続音声信号から音素，単語と文法を学習することが可能な新しい確率的生成モデルを提案する．提案手法では，各統計モデルのパラメータを相互に更新している．そのため，音素と単語と文法が相互に影響し合った学習を可能とする．実験では，文法学習を含む提案手法が，文法学習をしない従来手法よりも高い精度で連続音声信号を音素と単語に分割・分類できることを示した．また，文法学習が文中の単語数の正確な推定に大きく寄与することを示した．

1　はじめに

人間の幼児は正解を与えられなくとも，他者が発音した音声信号から言語を獲得することが可能である [1]-[3]．明確な境界やラベルがない状態で，連続音声信号を単語や音素に分割し，単語の遷移規則である文法を学ぶことは，言語学習において重要である．

音声認識の分野では，大規模なラベル付きデータやコーパスを利用した事前学習モデルなどの教師あり学習の手法が主に用いられている [4]-[7]．しかし，人間は音声信号から音素，単語，文法を学習する際に大量のラベル付きデータやコーパスを必要としないため，これらの手法は人間の言語学習とは異なる．そのため，ラベル付きデータセットやコーパスに依存せず，音声信号を音素と単語に分割・分類し，さらに文法を教師なしで学習できるモデルが求められている．

一方，音素と単語を教師なしで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 [8]．また，我々は Gaussian Process Hidden Semi Markov Model (GP-HSMM)[9][10] 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 [11]．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，文法学習は実現できていない．

そこで本稿では，二重分節構造を持つ連続音声信号から音素と単語だけでなく，文法も教師なしで学習するための確率的生成モデル（PGM）を提案する．提案モデルは GP-HSMM と HSMM で構成されている．GP-HSMM は連続音声を音素に分割する．HSMM は音素列を単語に分割し，単語クラスに分類することで，単語クラスの遷移として文法を学習する．また，提案手法では，GP-HSMM と HSMM をモジュール化し相互に学習する Symbol Emergence in Robotics tool KIT (Serket)[12][13] を使用し，音素と単語・文法を学習する．実験では，文法構造を含む日本語音声データセットを作成し，提案手法の音素，単語，文法を教師なしで学習する能力を検証した．また，文法学習によって得られた単語クラスが音素学習に影響を与え，推定精度を向上させることを示す．

2　提案手法

図 1 が提案手法のグラフィカルモデルである．この確率モデルは下位層が GP-HSMM，上位層が HSMM で構成されたモデルであり，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．

また，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，本稿では Serket のメッセージパッシング法によって各階層を交互に推論することで，モデル全体

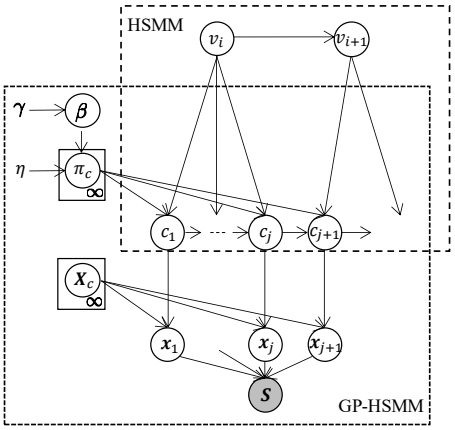

図 1　提案手法のグラフィカルモデル

のパラメータを最適化する．

2.1　生成過程

上位層の HSMM では，品詞に相当する単語クラス v_i が直前のクラス 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[11] とは異なる部分である．次に，単語クラス v_i から単語を構成する音素クラス c_j が，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．

$$c_j \sim P(c|c_{j-1}, \pi_c, v_i) \tag{2}$$

遷移確率 π_c の生成には階層ディリクレ過程 (HDP)[14] 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 [15]．

$$\beta \sim \mathrm{GEM}(\gamma) \tag{3}$$

$$\pi_c \sim \mathrm{DP}(\eta, \beta) \tag{4}$$

HDP 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 に対応する音声信号 $\boldsymbol{x}_j$ はパラメータ $\boldsymbol{X}_c$ を持つガウス過程から生成される．

$$\boldsymbol{x}_j \sim \mathcal{G}P(\boldsymbol{x}|\boldsymbol{X}_c) \tag{5}$$

ただし，$\boldsymbol{X}_c$ は音素クラス c に分類された音声信号の集合を表す．λ は音素の長さを決定するポアソン分布のパラメータである．観測された音声信号 S は，生成された $\boldsymbol{x}$ を連結することで得られると仮定する．データの連続性を柔軟に表現できる GP-HSMM を下位層に利用することで，音声のような連続信号を分節化することが可能である．

Algorithm 1 Mutual learning of GP-HSMM and HSMM

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

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

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

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

2.2　パラメータの推論

提案手法は GP-HSMM と HSMM で構成されており，モデルの階層が深く，パラメータの推定が困難である．そこで，モデルを構成する PGM をモジュール化し，モジュール間で相互にパラメータを更新する Serket を応用することによって全体に最適なパラメータを学習する．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クラス v から音素クラス c が生成される条件付き確率 $P(c|v)$ を計算し，下位層 (GP-HSMM) に送る．GP-HSMM では受け取った $P(c|v)$ 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を最適化する．

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 Forward 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 $x_{t-k:t}$ が音素クラス c となる前向き確率は次式となる．

$$\alpha_p[t][k][c] = \mathcal{G}P(\boldsymbol{x}_{t-k:t}|\boldsymbol{X}_c)P(c|v_i)P_{len}(k|\lambda_p) \times \sum_{k'=1}\sum_{c'=1} P(c|c', \pi_{c'})\alpha_p[t-k][k'][c'] \tag{6}$$

ただし，$P_{\mathrm{len}}(k|\lambda_p)$ 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，$P(c|c', \pi_{c'}, v_i) \approx AP(c|c', \pi_{c'})P(c|v_i)$ とした．A は正規化項である．$P(c|v_i)$ は，上位層で計算された単語クラス v_i から音素クラス c が発生する確率であり，この確率により文法構造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 Backward Sampling により音素クラス系列 C をサンプリングする．

$$k, c \sim \alpha[t][k][c]P(c_{j+1}|c) \tag{7}$$

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．Forward Filtering では，音素のタイムステップ j を終端とする長さ 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．

$$\alpha_w[j][k][v] = P(\boldsymbol{c}_{j-k:j}|v) \times \sum_{k'=1} \sum_{v'=1} P(v|v')\alpha_w[j-k][k'][v'] \tag{8}$$

ここで，$P(\boldsymbol{c}_{j-k:j}|v)$ は音素列 $\boldsymbol{c}_{j-k:j}$ が単語クラス v となる確率を表し，次のように計算される．

$$P(\boldsymbol{c}_{j-k:j}|v) = \frac{N_{v,w=\boldsymbol{c}_{j-k:j}} + \epsilon P_{\text{len}}(k|\lambda') \prod_{j'=j-k}^{j} P_{\text{phoneme}}(\boldsymbol{c}_{j'})}{N_v + \epsilon}, \tag{9}$$

$N_{v,w=\boldsymbol{c}_{j-k:j}}$ は単語クラス v において $\boldsymbol{c}_{j-k:j}$ が単語として出現した回数，N_v は単語クラス v 内の全単語数である．$P_{\text{len}}(k|\lambda')$ は単語長の確率である．また，ϵ は事前分布の重みである．

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 $\boldsymbol{S}$ から音素クラス列 $\boldsymbol{C}$ をサンプリング
2. 音素クラス列 $\boldsymbol{C}$ から単語列 $\boldsymbol{W}$ および単語クラス列 $\boldsymbol{V}$ をサンプリング
3. 各単語クラスから生成される各音素の確率 $P(c|v_i)$ を更新

3 実験

文法規則を持つ日本語音声データセット (KAKIKOKI データセット[1]) を作成し，提案手法を NPB-DAA と GP-HSMM-DAA と比較した．

3.1 実験設定

文法構造を含む KAKIKOKI データセットは，AIOI データセット[2]を参考に作成した．AIOI データセットは，日本語の母音 5 つで構成された 2 語または 3 語の発話音声から構成されているが，文法構造が含まれていない．KAKIKOKI データセットは，“aioi”, “aue”, “ao”, “ie”, “uo”, “kakikoki”, “kakuke”, “kako”, “kike”, “kuko”, “aka”, “iki” で構成される 2 語または 3 語の 75 文を，1 名の被験者が発話した音声信号データである．各単語は日本語の 10 種類の音節（音素）“a”, “i”, “u”, “e”, “o”, “ka”, “ki”, “ku”, “ke”, “ko” で構成されている．75 文のうち 25 文は 2 語文，50 文は 3 語文であり，文 G は以下の文法規則に従って決定される．

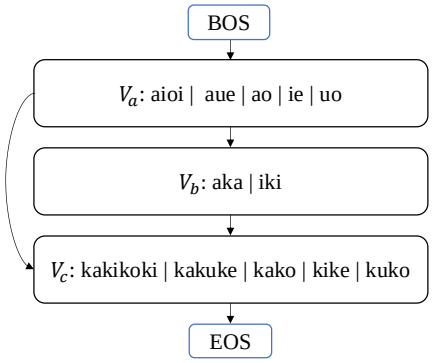

図 2 KAKIKOKI データセットの文法構造

- $G \rightarrow V_aV_bV_c \mid V_aV_c$
- $V_a \rightarrow$ aioi | aue | ao | ie | uo
- $V_b \rightarrow$ aka | iki
- $V_c \rightarrow$ kakikoki | kakuke | kako | kike | kuko

すなわち，3 語文の数は 50 通り（$5 \times 2 \times 5$），2 語文の数は 25 通り（5×5）で，全体の文の数は 75 通り（$50 + 25$）である．各文は 2 回ずつ読み上げられるため，KAKIKOKI データセットは合計 150 の音声信号で構成されている．さらに，評価のために各音声信号に単語と音節の正解ラベルを手作業で作成した．

特徴量は AIOI データセットと同様の方法で抽出した．音声信号のメル周波数ケプストラム係数（MFCC）を計算し，MFCC の 12 次元を Deep Sparse Autoencoder を用いて 6 次元に圧縮した．

提案手法を NPB-DAA と GP-HSMM-DAA と比較した．NPB-DAA は，音声信号の二重分節構造を利用して音素と単語を教師なしで学習するノンパラメトリック・ベイジアン手法である．この手法では，音声信号を音素単位に分割し，それらを統計的に単語にまとめることで，音声から自然な単語境界を抽出することが可能である．しかし，文法の学習は含まれていないため，音素や単語の学習に限定される．GP-HSMM-DAA は単語と音素の学習を行う手法であり，提案手法はこれを拡張して単語と音素に加えて文法も学習可能にした手法である．

本実験で使用した手法は初期パラメータの値に影響を受けるため、初期値を変えて 10 回学習を試行し，尤度が最大となる結果を評価に用いた．評価指標として，推定された単語または音素の系列と正解

1） https://github.com/naka-lab/kakikoki_dataset
2） https://github.com/EmergentSystemLabStudent/aioi_dataset

表 1　各手法の ARI の結果

手法	音節 ARI	単語 ARI
提案手法	0.409	0.614
GP-HSMM-DAA	0.304	0.391
NPB-DAA	0.359	0.489

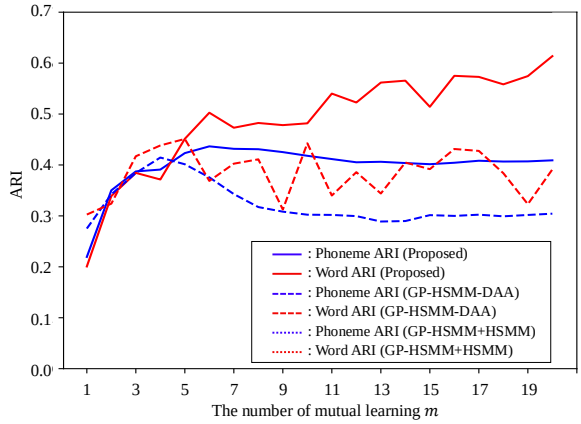

図 3　各手法における ARI の推移

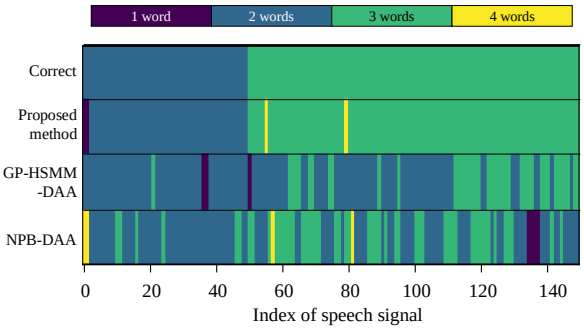

図 4　各音声データで推定された単語数

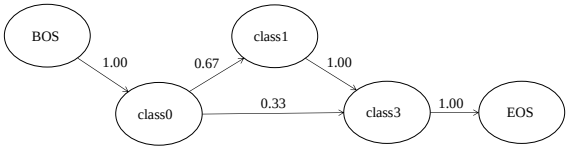

図 5　単語クラス間の遷移

との間の Adjusted Rand Index（ARI）[16] を用いた．ARI は，推定された系列が正解に近いほど 1 に近い値となる．

提案手法では単語クラス数を 4，相互学習の繰り返し回数を $M = 20$ として実験を行った．

3.2　実験結果

表 1 は各手法の ARI を示している．提案手法は音節および単語の精度で他の手法を上回った．NPB-DAA はガウス分布で音素をモデル化するため，音声信号の連続性を十分にモデル化できず，音節と単語の精度が提案手法に比べて低い結果となった．

図 3 は，提案手法と GP-HSMM-DAA の ARI の推移を示している．横軸は相互学習の回数，縦軸は ARI を表している．相互学習の回数が増加するにつれて提案手法の ARI が向上し，GP-HSMM-DAA を上回ったことが分かる．GP-HSMM-DAA では，単語クラスの遷移と単語クラスごとの音素の出現確率を十分にモデル化できていなかった．一方，提案手法では GP-HSMM の音素クラスの学習に HSMM の推定結果が適切に影響するため，音素クラスの推定誤りを減少させる結果になったと考えられる．また，HSMM では誤りの少ない音節が学習された結果、より正確な単語の学習が可能となった．

図 4 は，各音声信号において推定された単語数を示している．横軸は音声信号のインデックス，色は推定された単語数を表している．提案手法では，正しい単語数に最も近い推定結果が得られた．この結果からも，提案手法によって，文法構造の情報を効果的に活用できることがわかる．一方，GP-HSMM-DAA は文法を推定せず，文法情報を利用しないため，音声信号中の単語数を誤って推定する結果となった．

図 5 は，提案手法の各単語クラス間の遷移図を示している．図では，矢印上の数字はクラス間の遷移確率を表している．クラス数を 4 に設定したにも関わらず，全ての単語がクラス 0，1，3 に分類された．図 5 と図 2 を比較すると，推定されたクラス 0，1，3 がそれぞれ V_a，V_b，V_c に対応していることが確認できる．したがって，提案手法のパラメータにより適切な文法規則を学習できることを示した．

4　結論と今後の課題

本研究では，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．実験では文法の規則が存在するデータ (KAKIKOKI-データセット) 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データから自動的に推定可能な，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データセットであり，今後はより複雑な文法規則の学習が可能かを実データを用いた検証をする予定である．

謝辞

本研究は，JST ムーンショット型研究開発事業 JPMJMS2011 の支援を受けたものである．

参考文献

[1] Josiane Bertoncini. “Infants’ Perception of Speech Units: Primary Representation Capacities,” Dordrecht: Springer Netherlands, pp. 249–257, 1993.

[2] Linda Polka and Janet F. Werker. “Developmental Changes in Perception of Nonnative Vowel Contrasts,” Journal of Experimental Psychology: Human Perception and Performance, Vol. 20, No. 2, pp. 421, American Psychological Association, 1994.

[3] Yutaka Sato, Mahoko Kato, and Reiko Mazuka. “Development of Single/Geminate Obstruent Discrimination by Japanese Infants: Early Integration of Durational and Non-durational Cues,” Developmental Psychology, Vol. 48, No. 1, pp. 18, American Psychological Association, 2012.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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), pp. 4067–4074, 2018.

[11] Masatoshi Nagano and Tomoaki Nakamura. “Unsupervised Phoneme and Word Acquisition from Continuous Speech Based on a Hierarchical Probabilistic Generative Model,” Advanced Robotics, Vol. 30, No. 11–12, pp. 1253–1265, 2023.

[12] Tomoaki Nakamura, Takayuki Nagai, and Tadahiro Taniguchi. “Serket: An Architecture for Connecting Stochastic Models to Realize a Large-Scale Cognitive Model,” Frontiers in Neurorobotics, Vol. 12, pp. 25, Frontiers, 2018.

[13] Tadahiro Taniguchi, Tomoaki Nakamura, Masahiro Suzuki, Ryo Kuniyasu, Kaede Hayashi, Akira Taniguchi, Takato Horii, and Takayuki Nagai. “Neuro-SERKET: Development of Integrative Cognitive System through the Composition of Deep Probabilistic Generative Models,” New Generation Computing, pp. 1–26, Springer, 2020.

[14] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. “Hierarchical Dirichlet Processes,” Journal of the American Statistical Association, Vol. 101, No. 476, pp. 1566–1581, 2006.

[15] Jayaram Sethuraman. “A Constructive Definition of Dirichlet Priors,” Statistica Sinica, pp. 639–650, JSTOR, 1994.

[16] Lawrence Hubert and Arabie Phipps. “Comparing Partitions,” Journal of Classification, Vol. 2, pp. 193–218, 1985.