階層ディリクレ過程隠れ言語モデルと深層学習を用いた語彙獲得過程の計算論
A Computational Approach for Vocabulary Acquisition with Hierarchical Dirichlet Process-Hidden Language Model and Deep Learning.

林 楓 †，中島 諒 †，長坂 翔吾 †，谷口 忠大 ‡
Kaede Hayashi, Ryo Nakashima, Shogo Nagasaka, Tadahiro Taniguchi
† 立命館大学 情報理工学研究科, ‡ 立命館大学 情報理工学部
Graduate School of Information Science and Engineering, Ritsumeikan University,
Dept. of Information Science and Engineering, Ritsumeikan University.
k.hayashi@em.ci.ritsumei.ac.jp

Abstract

Human infants have been called 'stastical learners' capable of discovering words from speech signals. For a computational understanding of the language acquisition by infants, many researchers have developed machine learning models that can acquire knowledge about words and phonemes. The nonparametric Bayesian double articulation analyzer (NPB-DAA) is one of the unsupervised methods for a word discovery and can simultaneously learn a language model and an acoustic model. By integrating the NPB-DAA with the deep sparse autoencorder, it outperformed pre-existing unsupervised learning methods. In this paper, we show a new experimental result about word discovery from multiple Japanese speakers' speech signals. The results confirmed its effectiveness to the signle speaker data, but show the difficulty in the multiple-speaker data.

Keywords — Bayesian Nonparametrics, Deep Learning, Speech Recognition, Unsurpervised Learning, Word Discovery

1. はじめに

人間の幼児は身の回りから得られるセンサ情報を手がかりに言語を獲得していく．養育者の発音した音声信号から語彙を獲得するとき，幼児は時系列音声データからの単語分割問題を解いていることになる．つまり，幼児は言語を学習する過程で，言語に関する事前知識を持たずに時系列音声信号を認識し，単語列に分節化した上で，語彙を獲得する必要がある．このことから，単語分割および語彙発見は幼児が自律的に言語を獲得している過程の基盤となる技術であると言える．月齢 8ヶ月の幼児は音同士の統計的な共起性を用いて連続音声信号を分節化し，単語を抽出できることが示されている [1].

語彙獲得に関する既存研究として，事前の辞書なしでテキストデータを単語単位に分節化する先行研究 [2]，[3] がある．Mochihashi らはノンパラメトリックベイズに基づく nested Pitman-Yor language model (NPYLM) を提案し，高い精度の教師なし形態素解析を可能にした [2]．しかし，テキストデータからの語彙獲得と比べて，音声信号からの語彙獲得は困難である．この理由として，まず音声認識結果には音素認識誤りが頻繁に含まれることが挙げられる．また，初期の学習段階において，幼児は音素を完全に知らないままで語彙を獲得しなければならない．これに対して，Neibig らは NPYLM を拡張し，音素認識誤りを含んだ音声認識結果をラティスとして出力させたものを教師なし形態素解析し，その結果とともに音声認識結果を確定させる Latticelm を提案したが，この手法には音素獲得は含まれていない [3].

幼児の語彙獲得を再現するには音素獲得が必要であり，音素獲得と語彙獲得は相互に作用し合いながら行われると考えられている．これらの獲得過程の相互関係をモデル化することは，人間の言語獲得を計算論的に理解するための積年の課題となっている．近年，Lee らと Kamper らは音声信号からの教師なし語彙獲得手法を提案した [4]，[5]．また，Taniguchi らは，hierarchical Dirichlet process-hidden language model (HDP-HLM) とこれの潜在変数を推論する推論手法を構築することで，，人間の音声信号のみから言語モデルと音響モデルを同時に推論する機械学習手法 Nonparametric bayesian double articulation analyzer (NPB-DAA) を提案した [6]．HDP-HLM は音素獲得と語彙獲得を統合した確率的生成モデルであり，言語モデル (word model)

と音響モデル (acoustic model) を含んでいる．さらに，NPB-DAA に Deep sparse autoencorder (DSAE) を組み合わせることで，教師なし語彙獲得の精度向上を実現している [7]．この研究では，母音のみによって構成された単一話者による実音声データからの語彙獲得および音声認識において，既存の標準的な音声認識装置 [8] を上回る性能を持つことが示された．しかし，実際の幼児の言語獲得過程において，単一話者だけでなく複数の話者から語彙を獲得していると考えるのが妥当であるが，複数話者の音声信号からの語彙獲得は考慮されておらず，複数話者から得た音声信号に対して適用した際の有効性は検証されていなかった．したがって本研究では，NPB-DAA と DSAE の統合モデルを用いて，複数話者の実音声データから教師なし語彙獲得を行い，その実験結果について詳報する．

2. NPB-DAA

ここでは NPB-DAA の生成モデルである HDP-HLM について説明し，DSAE と NPB-DAA の統合モデルについて述べる．なお，NPB-DAA の推論過程などの詳細は [7] に記載されている．

2.1 HDP-HLM

HDP-HLM は，文を構成する Language model，単語を構成する Word model，音の情報を持つ Acoustic model からなる，HDP-HSMM [9] の拡張モデルである．図 1 に HDP-HLM のグラフィカルモデルを示す．z_s は HDP-HLM の super state であり，潜在単語を表している．また，i 番目の super state である $z_s = i$ は音素列 $w_i = (w_{i1}, ..., w_{ik}, ..., w_{iL_i})$ を持っている．ただし L_i は i 番目の単語 w_i の長さである．また，生成過程は以下のように表される．

$$
\begin{aligned}
\beta^{LM} &\sim \mathrm{GEM}(\gamma^{LM}) \\
\pi_i^{LM} &\sim \mathrm{DP}(\alpha^{LM}, \beta^{LM}) && i = 1,2,...,\infty \\
\beta^{WM} &\sim \mathrm{GEM}(\gamma^{WM}) \\
\pi_j^{WM} &\sim \mathrm{DP}(\alpha^{WM}, \beta^{WM}) && j = 1,2,...,\infty \\
w_{ik} &\sim \pi_{w_{ik-1}}^{WM} && i = 1,2,...,\infty \\
& && k = 1,2,...,L_i \\
(\theta_j, \omega_j) &\sim H \times G && j = 1,2,...,\infty \\
z_s &\sim \pi_{z_{s-1}}^{LM} && s = 1,2,...,S \\
l_{sk} &= w_{z_s k} && s = 1,2,...,S \\
& && k = 1,2,...,L_{z_s} \\
D_{sk} &\sim g(\omega_{l_{sk}}) && s = 1,2,...,S
\end{aligned}
$$

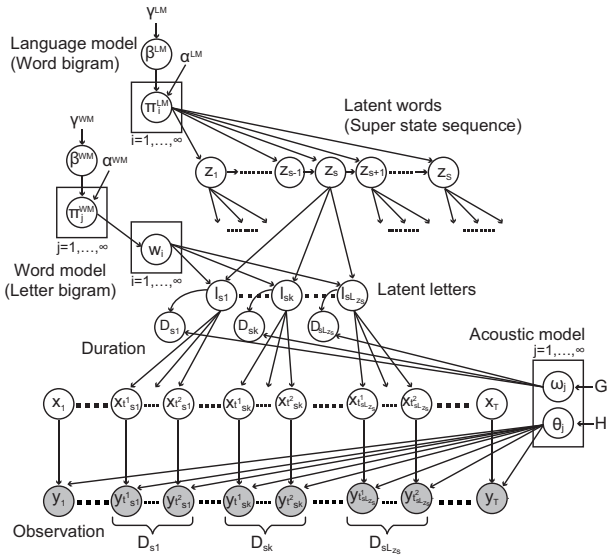

図 1　HDP-HLM のグラフィカルモデル

$$
\begin{aligned}
& && k = 1,2,...,L_{z_s} \\
x_t = l_{sk} & && t = t_{sk}^1, ..., t_{sk}^2 \\
& && t_{sk}^1 = \sum_{s'<s} D_{s'} + \sum_{k'<k} D_{sk'} + 1 \\
& && t_{sk}^2 = t_{sk}^1 + D_{sk} - 1 \\
y_t = h(\theta_{x_t}) & && t = 1,2,...,T
\end{aligned}
$$

ここで，β^{WM} は基底測度であり，α^{WM} と γ^{WM} は単語を生成する word model のハイパーパラメータである．$\mathrm{DP}(\alpha^{WM}, \beta^{WM})$ は潜在文字 j から次の文字への遷移確率 π_j^{WM} を出力する．β^{LM} は基底測度であり，α^{LM} と γ^{LM} は単語バイグラムを生成するハイパーパラメータである．また，$\mathrm{DP}(\alpha^{LM}, \beta^{LM})$ は潜在単語 i から次の単語への遷移確率 π_i^{LM} を出力する．また，上付きの文字 LM と WM はそれぞれ LM: Language Model と WM: Word Model を意味している．潜在単語 w_i に含まれる潜在文字は $\pi_{w_{ik-1}}^{WM}$ から順にサンプルされる．HDP-HLM では，持続分布は各潜在文字 l_{sk} によって決まる．潜在文字 l_{sk} の持続時間 D_{sk} は $g(\omega_{l_{sk}})$ から生成される．ただし，l_{sk} は s 番目の潜在単語 w_s の中の k 番目の潜在文字であり，$\omega_{l_{sk}}$ は潜在文字 l_{sk} の持続パラメータである．また，潜在単語 w_s の持続時間は $D_s = \sum_{k=1}^{L_{z_s}} D_{sk}$ となる．

HDP-HLM では潜在単語 z_s が潜在文字系列 $l_{sk} = w_{z_s k}$ $(k = 1,2,...,L_{z_s})$ を決定する．その後 w_{z_s} に基づいて，l_{sk} の持続時間 D_{sk} が生成され，出力分布 $h(\theta_{x_t})$ から観測データ y_t が生成される．これにより，潜在単語としてチャンク化される区間は一定の遷移パターンを持つデータとしてモデル化される．

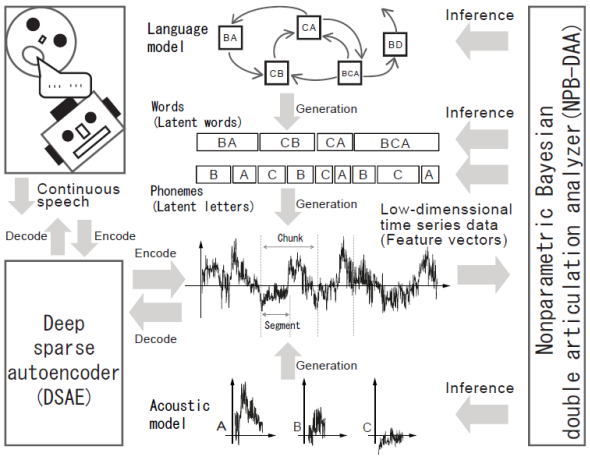

図 2　ロボットが実音声データから音素と語彙を獲得するための提案手法の全体像

2.2　NPB-DAA と DSAE の統合モデル

本研究で提案する機械学習手法の全体像を図 2 に示す．DSAE は実音声の信号から音声特徴を表現する時系列データを抽出している．そして，NPB-DAA はその時系列データを解析している．ここでの学習手法はすべて教師なし学習であり，音声信号からの特徴抽出や単語分割も教師なしで行われている．したがって，提案モデルである NPB-DAA と DSAE の統合モデルは，幼児が行う連続音声信号からの単語の発見や，音素の知識を得るための学習過程を説明するモデルの候補となりうる．

3.　実験

DSAE を組み合わせた NPB-DAA を用いて，複数人の実音声データから教師なし語彙獲得を行う．実験では 4 話者の実音声データから単一話者ごとに推定し評価平均をとったところ，DSAE を組み合わせた NPB-DAA の性能は 1 話者のみの実音声データを使用した実験 [7] での性能を上回った．しかし，4 話者の音声データから同時推定した場合の性能は，話者ごとに推定した場合の性能を大きく下回った．さらに，DSAE を用いて，4 話者のデータセットから抽出した特徴量の分布を確認したところ，男性話者と女性話者の特徴空間に大きなずれが生じていることがわかった．

3.1　実験条件

日本語を母国語とする男女各 2 人に，30 文を自然なスピードで 2 回ずつ繰り返し読んでもらい，合計 60 × 4 文分の音声を録音した．ただし，実験で用いた文は，日本語母音列 {a, i, u, e, o} のみから構成された 5 単語 {aioi, aue, ao, ie, uo} の組み合わせで作られたものである．30 文のうち，25 文は 2 語文（”aioi aue”, ”uo ie”など）でのり，5 文は 3 語文である．実験では，提案手法のある DSAE を組み合わせた NPB-DAA [7] と特徴ベクトルとして MFCC を用いた NPB-DAA の [6], conventional DAA [10]，そして既存の音声認識装置である Julius [8] を比較した．また，NPYLM と Latticelm のソフトウェア latticelm [11] を用いた．その他の実験条件については [7] と同様である．以降，2 名の女性話者データセットを K-DATA，M-DATA，2 名の男性話者のデータセットを H-DATA，N-DATA と呼称する．

3.2　実験結果

表 1 に 4 話者で平均をとった調整ランド指数（ARI）の値を示す．手法の末尾に（MAP）と表記のあるものは，他の手法が 20 試行した結果の平均をとっている一方，20 試行のうち事後確率が最大となった結果を示している．また，AM, LM の列にあるチェックマークは，それぞれ事前に音響モデルもしくは正しい単語辞書をもっている手法であるかどうかを示している．結果として，この単語分割タスクにおいて，提案手法は既存手法の性能を上回った．また，この実験では，4 話者分のデータセットから 1 話者ごとに推定し評価平均をとったが，それは 1 話者のみのデータを用いた実験 [7] で得られた評価値よりも良い結果を記録した．

表 2 は，DSAE を組み合わせた NPB-DAA を用いて，単一話者ごとに独立で解析した場合（one speaker）の平均 ARI と 4 話者分のデータセットから同時推定した場合（four speakers）の ARI の比較である．4 話者分のデータセットから同時に語彙獲得を行ったところ，単一話者ごとに独立で行った場合と比較して音素，単語ともに著しく性能が低下した．図 3 は 4 話者のデータセットから DSAE で抽出した 3 次元の特徴量を Principal component analysis (PCA) により 2 次元に変換した結果の特徴分布図である．そのうち，左図は 4 話者全員分の特徴分布図であり，右図は男性話者の H-DATA と女性話者の K-DATA 2 人分の特徴分布図である．ここから，男性話者と女性話者の特徴空間に大きなずれがあることがわかる．

4.　おわりに

本稿では，音声信号からの教師なし語彙獲得の手法として DSAE を組み合わせた NPB-DAA の紹介し，複数話者での実験により性能を評価した．今回の実験で，

表 1　推定された潜在文字と潜在単語の ARI

Method	Letter ARI	Word ARI	AM	LM
NPB-DAA with DSAE (MAP) [7]	**0.621**	**0.627**		
NPB-DAA with DSAE [7]	0.523	0.448		
NPB-DAA with MFCC (MAP) [6]	0.599	0.486		
NPB-DAA with MFCC [6]	0.561	0.394		
Conventional DAA with DSAE [10]	0.548	0.078		
Conventional DAA with MFCC [10]	0.591	0.091		
Julius (phoneme dictionary + NPYLM [2])	0.486	0.297	✓	
Julius (phoneme dictionary + latticelm [3])	0.538	0.319	✓	
Julius DNN (phoneme dictionary + NPYLM [2])	0.633	0.396	✓	
Julius (word dictionary)	0.586	0.486	✓	✓
Julius DNN (word dictionary)	0.674	0.769	✓	✓

表 2　単一話者ごとのデータから推定した場合の平均 ARI と 4 話者分のデータから同時推定した場合の ARI

NPB-DAA with DSAE	Letter ARI	Word ARI
one speaker (MAP)	**0.621**	**0.627**
one speaker	0.523	0.448
four speakers (MAP)	0.161	0.073
four speakers	0.234	0.139

DSAE を組み合わせた NPB-DAA が既存の標準的な音声認識装置よりも高い性能の語彙獲得が可能であることが示されたが，韻律的特徴や大規模語彙獲得を考えると解決すべき課題がいくつかある．この課題として話者依存性への対応が挙げられる．実験で，4 人分のデータを同時に DSAE と NPB-DAA にかけ解析した場合において，4 人分の音声データを別々に解析した場合の性能と比較して急激な性能低下がみられた．これは，話者による発声器官や発声方法などの違いにより，音声学上同じ音素と認識されるべきものが違う音素であると認識されているためだと考えられる．一方，人間の幼児は，様々な人から受け取った話者依存のある音声から語彙を獲得している．これを実現するには，顔認識などによって獲得した話者の識別情報を用いて，話者依存の音声から話者非依存な音声特徴を得るなどのアプローチが考えられるが，さらなる検討が必要である．さらに，幼児の語彙獲得を理解するためには，提案手法のような計算論的なモデルと実際の脳を照らし合わせて検討することは不可欠である．認知科学，神経科学の分野と協力して，人間の言語獲得の過程でこのような処理が脳のどの部位で行われているのか，また他の認知機能とどのように結びついているのかなどの議論を交わし，研究を進めていきたい．

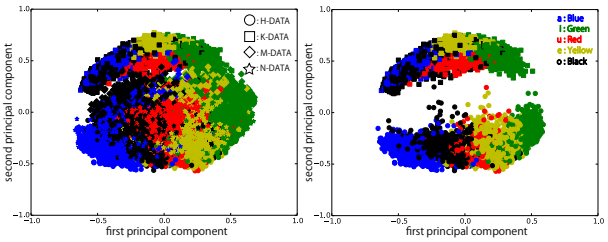

図 3　DSAE から PCA に変換した際の特徴分布図

参考文献

[1] Jenny R Saffran, Richard N Aslin, and Elissa L Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, 1996.

[2] Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 100–108. Association for Computational Linguistics, 2009.

[3] Graham Neubig, Masato Mimura, and Tatsuya Kawahara. Bayesian learning of a language model from continuous speech. *IEICE TRANSACTIONS on Information and Systems*, 95(2):614–625, 2012.

[4] Chia-ying Lee, Timothy J O'Donnell, and James Glass. Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics*, 3:389–403, 2015.

[5] Herman Kamper, Aren Jansen, and Sharon Goldwater. Fully unsupervised small-vocabulary speech recognition using a segmental bayesian model. In *Proc. Interspeech*, 2015.

[6] Tadahiro Taniguchi, Shogo Nagasaka, and Ryo Nakashima. Nonparametric bayesian double articulation analyzer for direct language acquisition from continuous speech signals. *IEEE Transactions on Cognitive and Developmental Systems*, 2016, accepted.

[7] 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*, 30(11-12):770–783, 2016.

[8] Open-source large vocablary csr engine julius. `http://julius.sourceforge.jp/`.

[9] Matthew J Johnson and Alan Willsky. The hierarchical dirichlet process hidden semi-markov model. *arXiv preprint arXiv:1203.3485*, 2012.

[10] Tadahiro Taniguchi and Shogo Nagasaka. Double articulation analyzer for unsegmented human motion using pitman-yor language model and infinite hidden markov model. In *System Integration (SII), 2011 IEEE/SICE International Symposium on*, pages 250–255. IEEE, 2011.

[11] latticelm. `http://http://www.phontron.com/latticelm/index-ja.html/`.