4D2-OS-37d-3

ノンパラメトリックベイス二重分節解析器の TIDIGITS コーパスへの適用

Application of Nonparametric Bayesian Double Articulation Analyzer to TIDIGITS Corpus

多田 裕貴 *1 Tada Yuuki　幸 優佑 *2 Miyuki yuusuke　林　楓 *1 Hayasi Kaede　萩原 良信 *2 Hagiwara Yoshinobu　谷口 忠大 *2 Taniguchi Tadahiro

*1 1 立命館大学情報理工学研究科
Graduate School of Information Science and Engineering, Ritsumeikan University.

*2 立命館大学情報理工学部
Information Science and Engineering, Ritsumeikan University.

Human infants can discover words directly from unsegmented speech signals without any labeled data. Nonparametric Bayesian double articulation analyzer (NPB-DAA) which is a unsupervised methods can discover words from unsegmented speech signals. NPB-DAA showed high performance in experiments from speech data by combining deep sparse autoencoder (DSAE) in a previous study. However, NPB-DAA in which words were discovered verified word acquisition from vowel only speech corpus. We experimented with the TIDIGITS corpus to find the method for extracting feature to adapt to consonants. Experimental results showed that mel-frequency cepstrum coefficients and its dynamic features were evaluated with the higher score than those of the other features. However, the log-likelihood of this feature and adjusted rand index (ARI) has a weak correlation. Therefore, we found that the features obtained from the DSAE constructed for each feature are the most suitable.

1. はじめに

幼児の語彙獲得の過程において，連続音声信号からの単語分割が重要な要素であると知られている．また，人間の幼児は月齢 8ヶ月の段階において，音声信号を単語ごとに分割ができる [Saffran 96]．Nonparametric Bayesian double articulation analyzer (NPB-DAA) は実音声データから直接に単語分割を行う機械学習モデルの一つである [Taniguchi 16a]．NPB-DAA の生成モデルは階層ディリクレ過程隠れ言語モデルであり，音声言語に潜む単語と音素の二層構造を表現している．加えて，谷口らは NPB-DAA による音声信号からの教師なし単語分割の性能向上を図るため，深層学習の一種である Deep sparse autoencoder (DSAE) を用いて，入力データを低次元化した特徴量を NPB-DAA の入力データとして扱った (NPB-DAA with DSAE と呼ぶ) ．図 1 は音声信号列から推論するまでの流れを表した NPB-DAA with DSAE の概要である．この NPB-DAA with DSAE は母音列のみで構成された音声データからの教師なし語彙獲得の実験で高いクラスタリング性能を示した [Taniguchi 16b].

一方で，幼児の語彙獲得を実現するためには，子音を含めた音声データにも対応する必要がある．そして，一般的に子音を含んだ音声を認識するための特徴量には，各時刻における特徴量である静的特徴量に加え，各時刻における静的特徴量の時間方向の変化量である動的特徴量が用いられる．これらのことから，先行研究では，静的特徴量のみを扱っていたが，子音を含めた音声データを考慮するためには動的特徴量が必要であると考える．その上，どのような動的特徴量を考慮すれば，単語の分割に貢献するかを調査しなければならない．

本研究では，単語分節化に貢献する動的特徴量を考慮した特徴抽出方法を検討することを目的とする．そのために，DSAE を組み合わせた NPB-DAA のモデルを用いて，子音を含んだ音声コーパスの単語分割実験を行う．子音を含んだ音声コーパスには，TIDIGITS コーパス *1 を用いる．この TIDIGITS コーパスを用いる理由は子音が含んでいるコーパスであると同時に，NPB-DAA が短期間で学習できる可能性がある単語の数で構成されているからである．また，TIDIGITS コーパスは他の教師なし単語分割手法の実験でも用いられており，他の手法 [Kamper 15] との比較が可能になることが挙げられる．

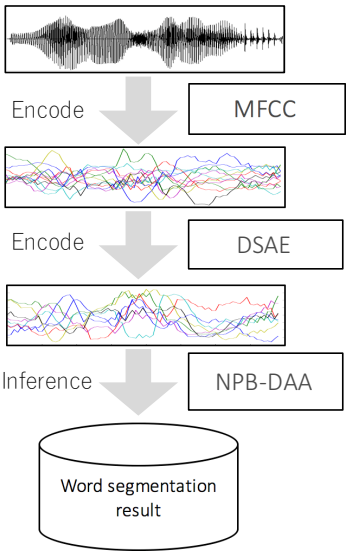

図 1: NPB-DAA with DSAE の概要図

連絡先: 多田 裕貴，立命館大学情報理工学研究科，525-8577，滋賀県草津市野路東，tada.yuuki@em.ci.ritsumei.ac.jp

*1 TIDIGITS:https://catalog.ldc.upenn.edu/ldc93s10

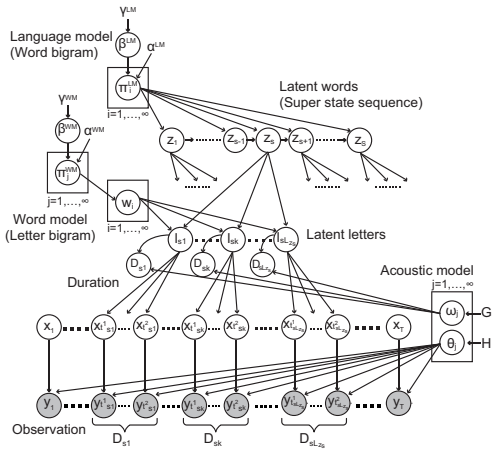

図 2: 階層ディリクレ隠れ言語モデルのグラフィカルモデル

2. NPB-DAA with DSAE

2.1 階層ディリクレ過程隠れ言語モデル

隠れディリクレ過程隠れ言語モデルは，単語の生成確率を保持している Language model，文字列の生成確率を保持している Word model，音素信号列の生成確率を保持している Acoustic model からなる生成モデルである．このモデルは Jhonson らが考案した HDP-HSMM を拡張したモデルである [Johnson 13]．階層ディリクレ隠れ言語モデルの生成過程を図 2 に示す．

$$\beta^{LM} \overset{iid}{\sim} \mathrm{GEM}(\gamma^{LM}) \tag{1}$$
$$\pi_i^{LM} \overset{iid}{\sim} \mathrm{DP}(\alpha^{LM}, \beta^{LM}) \tag{2}$$
$$\beta^{WM} \overset{iid}{\sim} \mathrm{GEM}(\gamma^{WM}) \tag{3}$$
$$\pi_j^{WM} \overset{iid}{\sim} \mathrm{DP}(\alpha^{WM}, \beta^{WM}) \tag{4}$$
$$w_{ik} \overset{iid}{\sim} \pi_{w_{ik-1}}^{WM} \tag{5}$$
$$(\theta_j, w_j) \overset{iid}{\sim} H \times G \tag{6}$$
$$z_s \overset{iid}{\sim} \pi_{z_{s-1}}^{LM} \tag{7}$$
$$l_{sk} = w_{z_s k} \tag{8}$$

$$D_{sk} \sim g(\omega_{l_{sk}}) \qquad s = 1, 2..., S \qquad k = 1, 2, ..., L_{z_s} \tag{9}$$
$$x_t = l_{sk} \qquad t = t_{sk}^1, ..., t_{sk}^2 \tag{10}$$
$$t_{sk}^1 = \sum_{s'<s} D_{s'} + \sum_{k'<k} D_{sk'} + 1 \qquad t_{sk}^2 + D_{sk} - 1$$
$$y_t = h(\theta_{x_t}) \qquad t = 1, 2....., T \tag{11}$$

生成モデルにおける潜在単語は super state Z_s に対応し，i 番目の super state $z_s = i$ は i 番目の潜在単語の音素列 $w_i = (w_{i1}, ..., w_{ik}, w_{iL_i})$ を持つ．GEM は SBP:stick breaking process，DP は Dirchlet Process を表している，また，LM，WM はそれぞれ言語モデル（LM），単語モデル（WM）を表している．β^{LM} は言語モデルのディリクレ過程の基底測度を表し，α^{LM} は集中度パラメータ，γ^{LM} の SBP のハイパラメータである．そして，DP(α^{LM}，β^{LM}) は潜在単語 i 番目から次の状態への遷移確率を表す π_i^{LM} を生成する．同様に, β^{WM} と α^{WM} は各 super state z に間連付けられた単語モデルの遷移確率分布 π_j^{WM} の基底測度と集中度パラメータであり，γ^{LM} は SBP のハイパーパラメータである．そして，DP(α^{WM}, β^{WM}) は潜在音素 j 番目から次の状態への遷移確率を表す単語モデル π_j^{WM} を生成する．i 番目の潜在単語 w_i が持つ潜在音素列は $\pi_{w_{ik-1}}^{WM}$ から順にサンプルされ i 番目の潜在単語における k 番目の潜在音素は w_{ik} で表せられる．音素信号を出力する出力分布 h と持続時間分布 g は，基底測度 H，G から生成される j 番目の潜在音素に関するパラメータ θ_j, ω_j をそれぞれ持つ．変数 z_s は潜在単語列における s 番目の単語であり，HDP-HSMM における super state に対応する．また，D_s は z_s のフレーム持続時間，$l_{sk} = w_{z_s k}$ は s 番目の潜在単語が持つ k 番目の潜在音素，そして，D_{sk} は l_{sk} のフレームで表せる．変数 y_t と x_t はフレーム単位の時間 t における観測データとその隠れ状態を表している．

次に，潜在単語列の s 番目の潜在単語 z_s が持つ k 番目の潜在音素 l_{sk} における持続時間 $D_{z_s k}$ は持続時間分布 $g(w_{l_{sk}})$ からサンプリングされる．ここで $g(w_{l_{sk}})$ は潜在音素がもつ持続時間に関するパラメータである．これらの結果から，潜在単語 w_{z_s} の持続時間は $D_s = \sum_{k=1}^{L_{z_s}} D_{sk}$ と定まる．持続時間分布 g がポアソン分布だと仮定すると，潜在単語 z_s もポアソン分布に従う．この場合の持続時間のポアソン分布パラメータは $\sum_{k=1}^{L_{z_s}} w_{l_{sk}}$ となる．HDP-HLM では潜在単語 z_s は潜在音素列 $l_{sk} = w_{z_s k} (k = 1, 2....., L_{z_s})$ から決定される．決定された順列 w_{z_s} に基づいて，l_{sk} の持続時間 D_{sk} はサンプリングされる．そして，観測データ y_t は $x_t = l_{s(t)k(t)}$ に対応する出力分布 $\mathrm{h}(\theta_{x_t})$ から生成される．これにより，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．詳しくは [Taniguchi 16a] を参考にされたい．

2.2 Deep sparse autoencoder (DSAE)

DSAE は Sparse autoencoder[Ng 11] を多層化したニューラルネットワーク構造を持つモデルである．特徴量を低次元化し，より抽象的な特徴量に変換させるために用いる．詳しくは [Taniguchi 16b] を参考にされたい。

3. 子音を含んだコーパスに対する適用実験

3.1 実験目的

本実験では，NPB-DAA を用いて，子音を含む音声コーパスからの教師なし語彙獲得の実験を行う．また，NPB-DAA に対して，動的特徴量の考慮する特徴量抽出方法を検討することを目的とする．

3.2 実験条件

音声データに TIDIGITS コーパスの女性話者 1 人分のデータを使用した．TIDIGITS コーパスとは，英語の発話音声データベースである．発話内容は 1 から 9 の数字と数字 0 の発話を 2 種類の合計 11 単語をランダムに並べた，英数字発話音声である．この音声は 77 文ある．また，無音区間も存在し，無音区間にも正解ラベルを与える．本実験では，NPB-DAA の入力する特徴量として，4 種類の動的特徴量の抽出方法を用い，評価する．4 つの動的特徴量を含む特徴抽出法を図 3 に示す．

1 つ目は音声データから 12 次元の Mel-Frequency Cepstrum Coefficients (MFCC) を抽出し，MFCC 及びその動的特徴量を NPB-DAA の入力とする方法である (MFCC+ΔMFCC+Δ^2MFCC とする) ．2 つ目に，1 つ目の特徴量抽出方法と同じく，音声データから MFCC 及

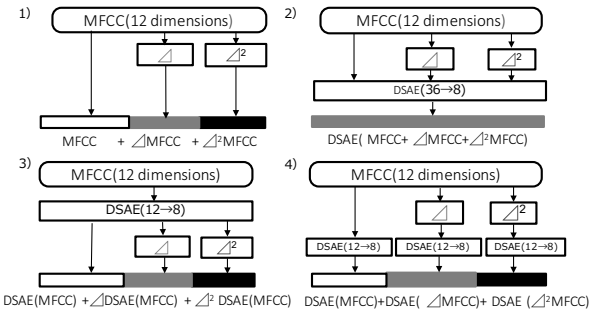

図 3: 動的特徴量を含む特徴抽出法

びその動的特徴量を取得し，DSAE で低次元化したものである (DSAE(MFCC+ΔMFCC+Δ^2MFCC) とする) ．3 つ目に，音声データから 12 次元の MFCC を抽出し，その特徴量を DSAE で 8 次元まで低次元化した特徴量及び，その動的特徴を NPB-DAA の入力とする方法である (DSAE(MFCC)+ΔDSAE(MFCC)+Δ^2DSAE(MFCC) とする) ．4 つ目に，音声データから 12 次元の MFCC を抽出し，MFCC 及びその動的特徴量をそれぞれ別の DSAE で低次元化し，それぞれ低次元化した特徴量を足し合わせた 24 次元の特徴量を用いる方法である (DSAE(MFCC)+DSAE(ΔMFCC)+DSAE(Δ^2MFCC) とする) ．比較対象として sphinx[*2] で音素認識を行い，sphinx の出力結果に対して，latticelm[Neubig 10] を使用して単語分割したものを用いる．

3.3 実験結果

本タスクは教師なし学習の課題であるので，時系列データのクラスタリングという視点から評価を行った．正解ラベルと NPB-DAA によって推定されたラベルとのクラスタリング性能を Adjusted rand index (ARI) を用いて，評価する [Hubert 85]．ARI はクラスタリング結果が正解ラベルと一致する場合は 1.0 に近い値を示し，正解ラベルと一致しない場合は 0.0 に近い値を示す．NPB-DAA を用いて，推定した評価結果を表 1 に示す．表 1 の ARI の値は 10 回のモデルの尤度が最大のときの推定結果の ARI 値 (MAP) とそれぞれの特徴量を用いて，10 回試行した平均値を採用している．この結果から，MFCC 及びその動的特徴量を加えた特徴量が 4 つの特徴量抽出方法の中では，平均と MAP どちらの評価値に対しても最も良い評価となった．

次に，モデルの尤度を高めることでより高い ARI の値を見込めるかを調査するために，4 つの手法に対しての 10 回の ARI とそれぞれの尤度の相関図を図 4 と図 5 に示した．この相関図の結果から，DSAE(MFCC+ΔMFCC+Δ^2MFCC) と DSAE(MFCC)+ DSAE(ΔMFCC)+DSAE(Δ^2MFCC) では，相関係数が 0.7 以上あり，モデルの尤度を高めることでクラスタリング性能が良くなることが見込める．

4. まとめ

本稿では，NPB-DAA に子音の含んだ音声コーパスである TIDIGITS コーパスを適用し，単語分割を行うために優位性のある特徴量の抽出方法の検証を行った．実験結果から DSAE を用いない MFCC のみの特徴量を用いた方がクラスタリング性能の評価値が高いことが分かった．しかし，DSAE を用いていない特徴量抽出方法と DSAE と DSAE の動的特徴量を用いた手法では，尤度と ARI に相関が見られず，モデルの尤度を上昇しても，クラスタリング性能の評価が上がることが保証されない．したがって，モデルの尤度と ARI の間に相関があり，尤度が最大のときに ARI の値が高い DSAE(MFCC)+DSAE(ΔMFCC)+DSAE(Δ^2MFCC) の特徴量を用いた方が妥当であると考える．今後の方針として，他の教師なし単語分割の手法とのクラスタリング精度評価の比較を行うことである．しかし，NPB-DAA におけるギブスサンプリングの計算量は TIDIGITS コーパスの実験では，77 音声の全 10047 フレームに対して，1 試行あたり 2CPU 構成の Xeon 14-Core で構成された PC で 1 試行あたり約 18 時間かかる． したがって，比較実験を行うためにも NPB-DAA の計算速度の効率改善も重要な課題である. こちらも同時に改善案を調査していく予定である．

*2 PocketSphinx 5prealpha : http://cmusphinx.sourceforge.net/

参考文献

[Saffran 96] Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport. "Statistical learning by 8-month-old infants." (1996).

[Taniguchi 16a] Tadahiro Taniguchi, Shogo Nagasaka, Ryo Nakashima Nonparametric Bayesian Double Articulation Analyzer for Direct Language Acquisition from Continuous Speech Signals, IEEE Transactions on Cognitive and Developmental Systems, Vol.8 (3), pp. 171-185 .(2016)DOI: 10.1109/TCDS.2016.2550591 (Open Access)

[Taniguchi 16b] Taniguchi, Tadahiro, et al. "Double articulation analyzer with deep sparse autoencoder for unsupervised word discovery from speech signals." Advanced Robotics 30.11-12 (2016): 770-783.

[Kamper 15] Kamper, Herman, Aren Jansen, and Sharon Goldwater. "Fully unsupervised small-vocabulary speech recognition using a segmental Bayesian model." Interspeech. 2015.

[Johnson 13] Johnson, Matthew J., and Alan S. Willsky. "Bayesian nonparametric hidden semi-Markov models." Journal of Machine Learning Research 14.Feb (2013): 673-701.

[Ng 11] Ng, Andrew. "Sparse autoencoder." CS294A Lecture notes 72.2011 (2011): 1-19.

[Hubert 85] Hubert, Lawrence, and Phipps Arabie. "Comparing partitions." Journal of classification 2.1 (1985): 193-218.

[Neubig 10] Neubig, Graham, et al. "Learning a language model from continuous speech." INTERSPEECH. 2010.

表 1: ARI による音素，単語ラベルのクラスタリング性能評価

Method	Feature		Letter ARI	Word ARI
NPB-DAA	MFCC+ΔMFCC+Δ^2MFCC	MAP	0.282	0.479
		平均	0.248	0.374
	DSAE(MFCC+ΔMFCC+Δ^2MFCC)	MAP	0.204	0.317
		平均	0.099	0.155
	DSAE(MFCC)+ΔDSAE(MFCC)+Δ^2DSAE(MFCC)	MAP	0.245	0.412
		平均	0.245	0.366
	DSAE(MFCC)+DSAE(ΔMFCC)+DSAE(Δ^2MFCC)	MAP	0.202	0.426
		平均	0.124	0.211
sphinx[*2] + latticelm [Neubig 10]	MFCC+ΔMFCC+Δ^2MFCC	平均	————	0.090

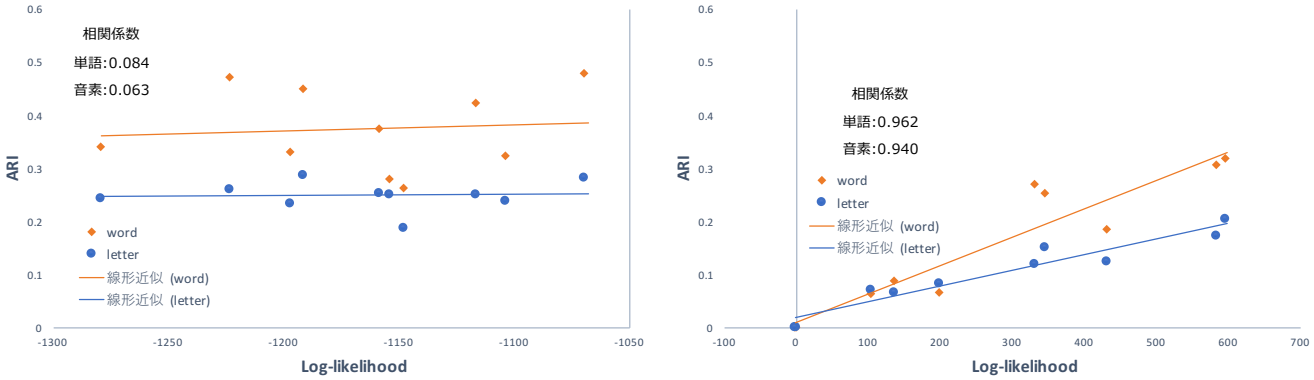

図 4: MFCC+ΔMFCC+Δ^2MFCC(左) と DSAE(MFCC+ΔMFCC+Δ^2MFCC)(右) の尤度と ARI の相関図

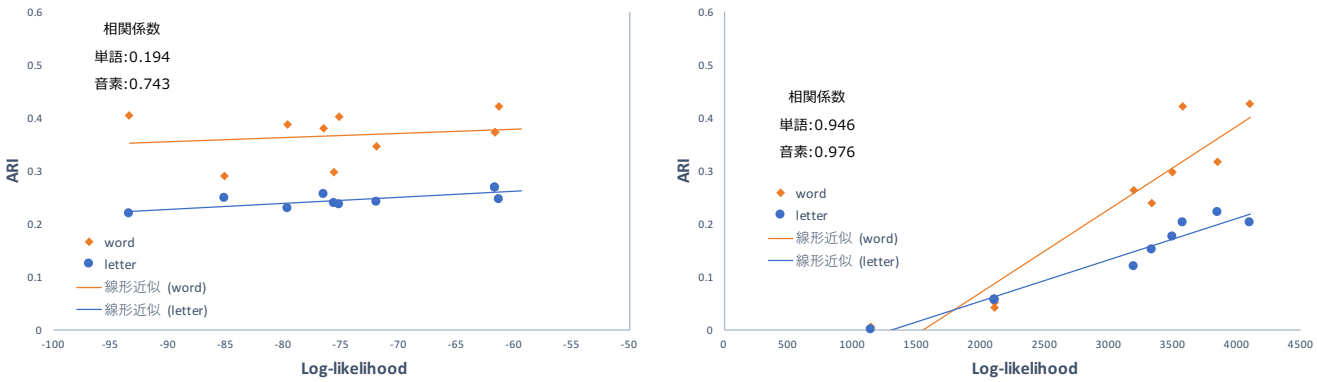

図 5: DSAE(MFCC)+ΔDSAE+Δ^2DSAE(左) と DSAE(MFCC)+ DSAE(ΔMFCC)+DSAE(Δ^2MFCC)(右) における尤度と ARI の相関図