The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

韻律情報を活用した二重分節解析器による自然音声からの語彙獲得

Double Articulation Analyzer with Prosody for Unsupervised Word Discovery

奥田 恭章 *1　尾崎 僚 *1　谷口 忠大 *1
Okuda Yasuaki　Ozaki Ryo　Taniguchi Tadahiro

*1立命館大学
Ritsumeikan University

Human infants discover words and phonemes using statistical information and prosody. For unsupervised word discovery, Taniguchi et al proposed the Nonparametric Bayesian Double Articulation Analyzer (NPB-DAA) which was able to segment speech data into word sequences. However, NPB-DAA uses only statistical information such as the mel-frequency cepstrum coefficients. In this paper, we extend NPB-DAA method using prosody, i.e., Prosodic DAA, for unsupervised word discovery. We use the second order differential of the fundamental frequency and the duration of silent as the prosody. We show in an experiment that Prosodic DAA outperforms NPB-DAA.

1. はじめに

人間の幼児は月齢 8ヵ月において，音声信号の分布情報や韻律情報 (Prosody，プロソディ) から，単語の単位を見いだせることが知られている．本研究では，教師なし学習によりプロソディを含む自然音声信号から，直接音響モデルと言語モデルを同時に推定することで，語彙獲得を行う手法を提案する．

音声信号から教師なし語彙獲得を行う手法にノンパラメトリックベイズ二重分節解析器 (Nonparametric Bayesian Double Articulation Analyzer: NPB-DAA) がある．しかしながら，現在の NPB-DAA ではプロソディによる音響特徴量の変化を扱うことができない．そこで本稿では NPB-DAA にプロソディを扱うような出力分布を新たに加えた Prosodic DAA を提案する．

また，プロソディを強調した日本語自然発話からなる実音声発話から教師なし語彙獲得する実験を行う．実験結果において，プロソディを用いて単語推定を行うことで NPB-DAA の音素及び単語推定性能が向上することを示す．

2. 先行研究

人間の語彙獲得過程において，月齢 8ヵ月の幼児は音声を単語ごとに分割することができる [Saffran 96]．また，幼児は語彙獲得の際に，音声学的性質であるプロソディを用いて音声を単語ごとに分割していることが知られている [Jusczyk 92]．このことから，幼児は音声信号の分布，音素と単語の共起性やプロソディを用いて語彙獲得をしていると考えられる．

谷口らは，二重分節構造を持つ時系列データの生成モデルとそのパラメータの推論手法を組み合わせて，NPB-DAA を提案した [Taniguchi 16a]．NPB-DAA は二重分節構造を持つ時系列データの解析手法である．二重分節構造は，単独では意味を持たない音素と，音素の組み合わせによって意味を持つ単語による二重の分節構造である．先行研究において，谷口らは母音のみの音声発話データからの単語分割を実現した [Taniguchi 16a]．また，音声特徴量を Deep Sparse Autoencoder (DSAE) を用いて，より高次の特徴量に変換することで単語分割精度が向上することを示した [Taniguchi 16b]．夛田らは子音を含んだ実音声発話データからの単語分割を行うために動的特徴量を用いた [Tada 17]．尾崎らは NPB-DAA にルックアップテーブルを導入することで計算量オーダを 3 次から 2 次に削減し，計算時間を 90%削減した [Ozaki 18]．

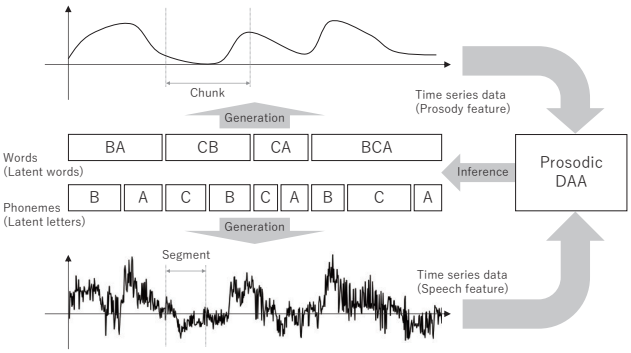

図 1: 本研究の概要図．二重分節解析器にプロソディを活用した単語分割を組み込む．ここで，Chunk は単語の継続長，Segment は音素の継続長をそれぞれ表す．

NPB-DAA では，一つの単語は同一の left-to-right の隠れセミマルコフモデル (Hidden semi-Markov Model: HSMM) [Johnson 13] に従うとしており，プロソディによる音声波形の変化は異なる単語として認識される．先行研究では，音響特徴量として，プロソディを含まない特徴量であるメル周波数ケプストラム係数 (Mel-Frequency Cepstrum Coefficients: MFCC) を用いており，プロソディによる音声波形の変化を考慮していない．そのため，幼児が語彙獲得過程で行うとされているプロソディによる単語分割を表現していない．そこで，本研究では NPB-DAA にプロソディを用いた単語分割を組み込むことを目的とする．本研究の概要図を図 1 に示す．

連絡先: 奥田 恭章，立命館大学 情報理工学研究科，滋賀県 草津市 野路東 1-1-1，okuda.yasuaki@em.ci.ristumei.ac.jp

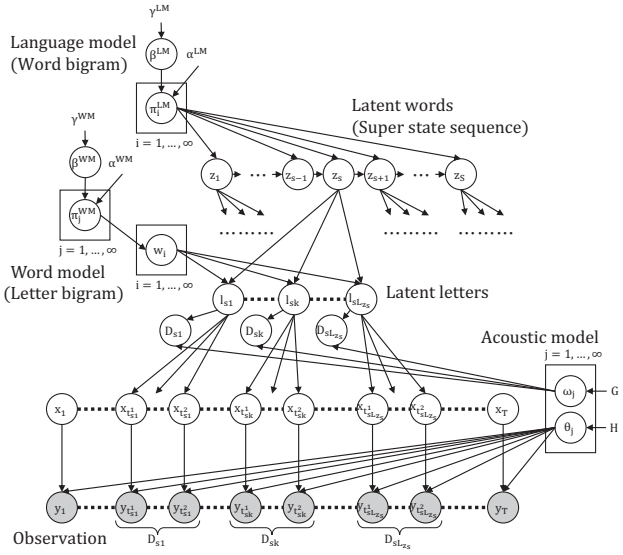

図 2: HDP-HLM のグラフィカルモデル [Taniguchi 16a]．このモデルは，二重分節構造を持つ時系列データの生成モデルである．

3. NPB-DAA

3.1 階層ディリクレ過程隠れ言語モデル

階層ディリクレ過程隠れ言語モデル (Hierarchical Dirichlet Process Hidden Language Model: HDP-HLM) は二重分節構造を持つ時系列データの生成モデルであり，これは階層ディリクレ過程隠れセミマルコフモデル (Hierarchical Dirichlet Process Hidden semi-Markov Model: HDP-HSMM) [Johnson 13] を二重分節構造の表現ができるように拡張したものである．図 2 に HDP-HLM のグラフィカルモデルを示す．その他，生成過程などについては [Taniguchi 16a] を参照されたい．

3.2 潜在単語のサンプリング

HDP-HLM の潜在単語 z_s は HDP-HSMM の推論手法で用いる backward filtering forward sampling の手続きを拡張してサンプリングされる．HDP-HLM の潜在単語 $z_s = i$ の backward message は以下の式で求めることが出来る．ここで，F_t は潜在単語の境界かどうかを表し，$F_t = 1$ であれば時刻 $t+1$ で単語が切り替わるとする．$B_t(i)$ は潜在単語 $z_{s(t)} = i$ が時刻 $t+1$ において，異なる潜在単語に遷移する尤度を表す．$B^*_t(i)$ は時刻 $t+1$ から潜在単語が $z_{s(t)} = i$ となる尤度を表す．

$$
\begin{aligned}
B_t(i) &:= p(y_{t+1:T}|z_{s(t)} = i, F_t = 1) \\
&= \sum_j B^*_t(j) p(z_{s(t+1)} = j|z_t = i) \qquad (1) \\
B^*_t(i) &:= p(y_{t+1:T}|z_{s(t+1)} = i, F_t = 1) \\
&= \sum_{d=1}^{T-t} B_{t+d}(i) p(D_{t+1} = d|z_{s(t+1)} = i) \\
&\quad \cdot p(y_{t+1:t+d}|z_{s(t+1)} = i, D_{t+1} = d) \qquad (2) \\
B_T(i) &:= 1 \qquad (3)
\end{aligned}
$$

潜在単語 $z_{s(t)} = i$ の潜在音素列 w_i を $w_i = (l_1, \cdots, l_{L_i})$ と置くと，$p(y_{t+1:t+d}|z_{s(t+1)} = i, D_{t+1} = d)$ は以下の式のように計算できる．ここで，$R^{(L_i,d)}$ は L_i 次元で要素の和が d となるような自然数ベクトル $\mathbf{r}$ の集合を表す．

$$
\begin{aligned}
p(y_{t+1:t+d}|i, d) &= \sum_{r \in R^{(L_i,d)}} \prod_{k=1}^{L_i} p(r_k|l_k) \\
&\quad \cdot \prod_{m=1}^{r_k} p(y_{t+m+\sum_{k'=1}^{k-1} r_{k'}}|l_k) \qquad (4) \\
R^{(L_i,d)} &= \left\{ r \in \{1, 2, \ldots\}^{L_i} | \sum_{k=1}^{L_i} r_k = d \right\} \qquad (5)
\end{aligned}
$$

式 4 は動的計画法を用いて効率的に計算することができる．時刻 t において潜在音素の遷移が発生し，その潜在音素が潜在単語中の k 番目の要素である確率を forward message と定義し，以下の式で再帰的に求めることができる．

$$
\begin{aligned}
\alpha_t(k) &= \sum_{d'=1}^{t-k+1} \alpha_{t-d'}(k-1) p(d'|l_k) \prod_{t'=1}^{d'} p(y_{t-t'+1}|l_k) \\
&\qquad (6) \\
\alpha_0(0) &= 1 \qquad (7)
\end{aligned}
$$

以上により，HDP-HLM において各時刻 t における．$B_t(i)$, $B^*_t(i)$ が求まるため，以下の式に従い潜在単語 $z_{s(t+1)}$ とその継続時間 $D_{s(t+1)}$ をサンプリングできる．ここで，$D^{\mathrm{sum}}_s = \sum_{s'<s} D_{s'}$ である．

$$
\begin{aligned}
&p(z_s = i|y_{1:T}, z_{s-1} = j, F_{D^{\mathrm{sum}}_{1:s}} = 1) = \\
&\qquad p(i|j) \beta_{D^{\mathrm{sum}}_{1:s}}(i) p(y_{D^{\mathrm{sum}}_{1:s}}|i) \qquad (8)
\end{aligned}
$$

$$
\begin{aligned}
&p(D_s = d|y_{1:T}, z_s = i, F_{D^{\mathrm{sum}}_{1:s}} = 1) = \\
&\quad p(y_{D^{\mathrm{sum}}_{1:s}+1:D^{\mathrm{sum}}_{1:s}+d}|d, i, F_{D^{\mathrm{sum}}_{1:s}} = 1) p(d) \frac{\beta_{D^{\mathrm{sum}}_{1:s}+d}(i)}{\beta^*_{D^{\mathrm{sum}}_{1:s}}(i)} \qquad (9)
\end{aligned}
$$

4. Prosodic DAA

本稿では，プロソディとして基本周波数 F_0 の二次微分及び無音区間の継続長の 2 つを用いる．無音区間の抽出では，ある閾値以下の音量が一定時間連続している区間を無音区間とし，その継続長を用いる．

次に，プロソディを HDP-HLM に組み込む方法を述べる．本稿では，プロソディはそれぞれ観測 y_t と同じフレーム数を持ち，単語の境界を表すインジケーター変数 F_t から生成されるものと仮定している．また，F_0 の二次微分を表す観測を $y^{\mathrm{F}_0}_t$ とし，無音区間長を表す観測を y^{sil}_t とする．プロソディを含む HDP-HLM のグラフィカルモデルを図 3 に示す．また，生成過程は以下のようになる．ここで GEM は Stick-Breaking Process を表し，DP は Dirichlet Process を表す．LM,WM はそれぞれ言語モデルと単語モデルを表している．また，β^{WM} は単語モデルにおける Dirichlet Process の基底測度を表し，α^{WM} と γ^{WM} はそれぞれ Dirichlet Process, Stick-Breaking Process のハイパーパラメータである．そして π^{WM}_i は潜在音素 i から次の状態への遷移確率を表している．同様に β^{LM} は言語モデルにおける Dirichlet Process の基底測度を表し，α^{LM} と γ^{LM} はそれぞれ Dirichlet Process, Stick-Breaking Process のハイパーパラメータである．そして π^{LM}_i は潜在単語 i から

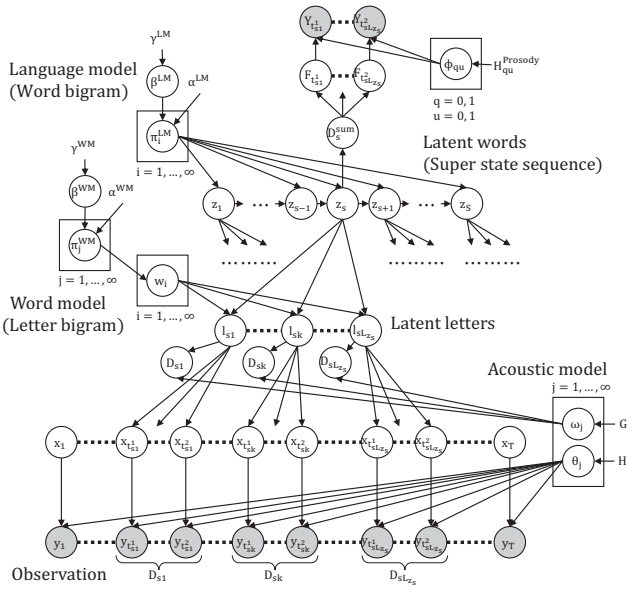

図 3: プロソディを含む HDP-HLM のグラフィカルモデル．潜在単語 z_s から z_s の継続長 D_s^{sum} が生成される．D_s^{sum} から単語境界を表す F_t が生成され，F_t から Y_t が生成される．

次の状態への遷移確率を表している．また，$Y_{1:T}$ は $y_{1:T}^{\mathrm{F_0}}$，$y_{1:T}^{sil}$ の集合を表し，$p(Y_t)\quad(t=1:T)$ は $p(y_t^{\mathrm{F_0}},\ y_t^{sil})$ である．

$$\beta^{\mathrm{LM}} \sim \mathrm{GEM}(\gamma^{\mathrm{LM}}) \tag{10}$$
$$\pi_i^{\mathrm{LM}} \sim \mathrm{DP}(\alpha^{\mathrm{LM}}, \beta^{\mathrm{LM}}) \quad (\mathrm{i}=1,\ldots) \tag{11}$$
$$\beta^{\mathrm{WM}} \sim \mathrm{GEM}(\gamma^{\mathrm{WM}}) \tag{12}$$
$$\pi_i^{\mathrm{WM}} \sim \mathrm{DP}(\alpha^{\mathrm{WM}}, \beta^{\mathrm{WM}}) \quad (\mathrm{i}=1,\ldots) \tag{13}$$
$$w_{ik} \sim \pi_{w_{ik-1}}^{\mathrm{WM}} \quad (i=1,\ldots)\quad (k=1,\ldots,L_i) \tag{14}$$
$$(\theta_j, \omega_j) \sim H \times G \quad (j=1,\ldots) \tag{15}$$
$$\phi_{qu} \sim H_{qu}^{Prosody} \quad (q=0,1)\quad (u=0,1) \tag{16}$$
$$z_s \sim \pi_{z_{s-1}}^{\mathrm{LM}} \quad (s=1,\ldots,S) \tag{17}$$
$$l_{sk} = w_{w_{z_s}k} \quad (s=1,\ldots,S)\quad (k=1,\ldots,L_{z_s}) \tag{18}$$
$$D_{sk} \sim g(\omega_{l_{sk}}) \quad (s=1,\ldots,S)\quad (k=1,\ldots,L_{z_s}) \tag{19}$$
$$x_t = l_{sk} \quad (t=t_{sk}^1,\ldots,t_{sk}^2) \tag{20}$$
$$t_{sk}^1 = \sum_{s'<s} D_{s'} + \sum_{k'<k} D_{sk'} + 1$$
$$t_{sk}^2 = t_{sk}^1 + D_{sk} - 1 \tag{21}$$
$$D_s^{sum} = \sum_{t=s1}^{sL_{z_s}} D_t \tag{22}$$
$$y_t \sim h(\theta_{x_t}) \quad (t=1,\ldots,T) \tag{23}$$
$$F_t = \begin{cases} 0 & (t=t_{s1}^1 : t'-1) \\ 1 & (t=t') \end{cases} \qquad (t'=t_{sL_{z_s}}^2) \tag{24}$$
$$y_t^{\mathrm{F_0}} \sim h^{\mathrm{F_0}}(\phi_{F_t 0}) \tag{25}$$
$$y_t^{sil} \sim h^{sil}(\phi_{F_t 1}) \tag{26}$$

また，プロソディによる単語分割を含む，潜在単語 $z_s=i$ の backward message は以下の式で求められる．

$$\begin{aligned}\beta_t(i) &:= p(y_{t+1:T}, Y_{t+1:T}|z_{s(t)}=i, F_t=1)\\ &= \sum_j \beta_t^*(j)p(z_{s(t+1)}=j|z_t=i) \qquad (27)\\ \beta_t^*(i) &:= p(y_{t+1:T}, Y_{t+1:T}|z_{s(t+1)}=i, F_t=1)\\ &= \sum_{d=1}^{T-t} \beta_{t+d}(i)p(D_{t+1}=d|z_{s(t+1)}=i)\\ &\quad \cdot p(y_{t+1:t+d}, Y_{t+1:t+d}|i,d) \qquad (28)\\ \beta_T(i) &:= 1 \qquad (29)\end{aligned}$$

潜在単語 $z_{s(t)}=i$ の潜在音素列 w_i を $W_i=(l_1,\cdots,l_{L_i})$ と置くと，$p(y_{t+1:t+d}, Y_{t+1:t+d}|z_{s(t+1)}=i, D_{t+1}=d)$ は以下の式のように計算できる．

$$p(y_{t+1:t+d}, Y_{t+1:t+d}|i,d) = \sum_{r\in R^{(L_i,d)}} \prod_{k=1}^{L_i} p(r_k|l_k)$$
$$\cdot \prod_{m=1}^{r_k} p(y_{t+m+\sum_{k'=1}^{k-1} r_{k'}}|l_k) p(Y_{t+m+\sum_{k'=1}^{k-1} r_{k'}}|l_k) \tag{30}$$

$$R^{(L_i,d)} = \{r\in\{1,2,\ldots\}^{L_i} | \sum_{k=1}^{L_i} r_k = d\} \tag{31}$$

forward message は，以下の式で再帰的に求めることができる．

$$\begin{aligned}\alpha_t(k) &= \sum_{d'=1}^{t-k+1} \alpha_{t-d'}(k-1)p(d'|l_k)\\ &\quad \cdot \prod_{t'=1}^{d'} p(y_{t-t'+1}, Y_{t-t'+1}|l_k) \qquad (32)\\ \alpha_0(0) &= 1 \qquad (33)\end{aligned}$$

以上により，HDP-HLM において各時刻 t における．$\beta_t(i)$，$\beta_t^*(i)$ が求まるため，以下の式に従い潜在単語 $z_{s(t+1)}$ とその継続時間 $D_{s(t+1)}$ をサンプリングできる．ここで，$D_{1:s}^{\mathrm{sum}} = \sum_{s'<s} D_{s'}$ である．

$$p(z_s=i|y_{1:T}, Y_{1:T}, z_{s-1}=j, F_{D_{1:s}^{\mathrm{sum}}}=1) =$$
$$p(i|j)\beta_{D_{1:s}^{\mathrm{sum}}}(i)p(y_{D_{1:s}^{\mathrm{sum}}}, Y_{D_{1:s}^{\mathrm{sum}}}|i) \tag{34}$$

$$p(D_s=d|y_{1:T}, Y_{1:T}, z_s=i, F_{D_{1:s}^{\mathrm{sum}}}=1) =$$
$$p(y_{D_{1:s}^{\mathrm{sum}}+1:D_{1:s}^{\mathrm{sum}}+d}, Y_{D_{1:s}^{\mathrm{sum}}+1:D_{1:s}^{\mathrm{sum}}+d}|d,i,F_{D_{1:s}^{\mathrm{sum}}}=1)$$
$$\cdot p(d)\frac{\beta_{D_{1:s}^{\mathrm{sum}}+d}(i)}{\beta_{D_{1:s}^{\mathrm{sum}}}^*(i)} \tag{35}$$

以上により，プロソディを考慮した二重分節構造の生成過程から事後分布のサンプリング式を導出した．

5. 実験

5.1 実験目的

本実験では，提案手法が自然発話からの語彙獲得が可能であるかを評価するために，プロソディを含む自然発話データセットを用いて，NPB-DAA と，提案手法 (以降 Prosodic DAA と呼称する．) の音素，及び単語分割精度を比較する．

5.2 実験設定

本実験では，日本語自然発話データセットを用いる．ここでは，自然発話を日常会話と相違ない発話と定義する．このデータセットは，日本語自然発話で構成された 70 文を 1 回ずつ男性の話者に発話してもらい収録した音声データセットである．本実験で用いる特徴量は，フレーム幅を 25[msec]，フレームシフト長を 10[msec] として変換された 12 次元の MFCC，MFCC の一次微分，及び二次微分のそれぞれの特徴量を用いる．また，それぞれ DSAE で 8 次元，5 次元，3 次元と段階的に抽出したものを結合した，合計 9 次元の特徴量を用いる．F_0 の推定には動的計画法を用いて F_0 を系列として推定する手法である Robust Epoch And Pitch EstimatoR[*1] を用いる．

5.3 実験条件

実験条件として，HDP-HLM の隠れ言語モデルのハイパーパラメータは，$\alpha^{\mathrm{LM}} = 10.0$ と $\gamma^{\mathrm{LM}} = 10.0$ とし，weak-limit 近似における単語上限数を 35 個とした．同様に，隠れ単語モデルでは，$\alpha^{\mathrm{WM}} = 10.0$ と $\gamma^{\mathrm{WM}} = 10.0$ とし，weak-limit 近似における音素上限数を 30 個とした．持続時間分布には，$\alpha_0 = 200$，$\beta_0 = 10$ のポアソン分布を仮定し，MFCC の出力分布には事前分布に $\mu_0 = 0$，Σ_0 に単位行列，$\kappa_0 = 0,01$，$\nu_0 = (\mathrm{dimension} + 2)$ の正規逆ウィシャート分布を持つ多変量ガウス分布と仮定した．さらに，提案手法では，プロソディの出力分布には事前分布に，$F_t = 0$ のとき $\mu_0 = 0$，Σ_0 に単位行列，$\kappa_0 = 100$，$\nu_0 = (\mathrm{dimension} + 2)$，$F_t = 1$ のとき $\mu_0 = 1$，Σ_0 に単位行列，$\kappa_0 = 2$，$\nu_0 = (\mathrm{dimension} + 2)$ の正規逆ウィシャート分布を持つ多変量ガウス分布と仮定した．上記条件において，100 イテレーションを 1 試行とし，これを 20 試行行う．

5.4 実験結果

本実験の結果として，音声データの各フレームに正解ラベルを付与し，クラスタリング性能を評価する．クラスタリング性能の定量的評価指標として，各試行で得られた最終サンプリング結果から，音素及び単語の調整ランド指数 (Adjusted Rand Index: ARI) を算出する．ARI はクラスタリング結果が正解ラベルと一致するとき 1 を取り，ランダムの場合は 0 をとる．

表 1，及び図 4 に Prosodic DAA，及び NPB-DAA のそれぞれの音素及び単語の ARI の平均を示す．表 1 より，2 つの手法の結果をみると，Prosodic DAA が音素 ARI，単語 ARI 共に大幅に高い値を出していることがわかる．また，Prosodic DAA と，NPB-DAA 間で t 検定を行ったところ，音素 ARI と単語 ARI 共に優位水準 1%で有意差が認められた．

6. まとめ

本稿では，NPB-DAA によるプロソディを考慮した単語分割を行う手法を提案した．実験では，プロソディを強調した日本語自然発話から語彙獲得実験を行い，提案手法が既存手法と比較して，高い精度で語彙獲得が可能であることを示した．

表 1: 音素及び単語分割の精度比較.

手法	音素 ARI	単語 ARI
Prosodic DAA	**0.370±0.022**	**0.671±0.054**
NPB-DAA	0.261±0.014	0.497±0.072

*1 REAPER: Robust Epoch And Pitch EstimatoR : https://github.com/google/REAPER

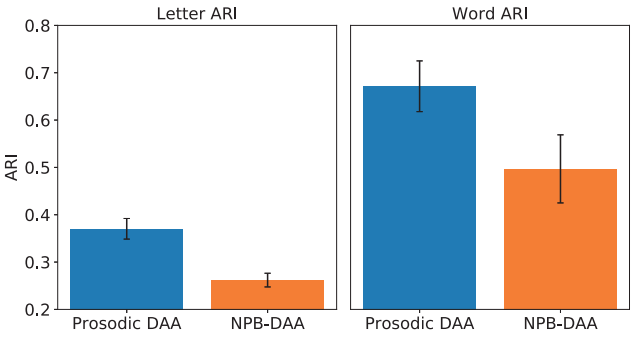

図 4: 音素及び単語分割の精度比較.

今後の課題として，プロソディを強調していない，日本語自然発話からの語彙獲得を行うことが考えられる．

参考文献

[Saffran 96] Saffran, Jenny R., Aslin, Richard N. and Newport, ElissaL.: Statistical Learning by 8-Month-Old Infants, American Association for the Advancement of Science, Vol.274, No.5294, pp.19261928, 1996

[Jusczyk 92] Jusczyk, Peter W and Hirsh-Pasek, Kathy and Nelson, Deborah G Kemler and Kennedy, Lori J and Woodward, Amanda and Piwoz, Julie.:Perception of acoustic correlates of major phrasal units by young infants, Cognitive psychology, Vol.24, Num.2, pp.252-293, 1992

[Taniguchi 16a] Taniguchi, Tadahiro and Nagasaka, Shogo and Nakashima, Ryo.: Nonparametric Bayesian Double Articulation Analyzer for Direct Language Acquisition From Continuous Speech Signals, IEEE Trans. Cognitive and Developmental Systems, Vol.8, Num.3, pp.171185, 2016

[Taniguchi 16b] Taniguchi, Tadahiro and Nakashima, Ryo and Liu, Hailong and Nagasaka, Shogo.: Double articulation analyzer with deep sparse autoencoder for unsupervised word discovery from speech signals, Advanced Robotics, Vol.30, Num.11-12, pp.770783, 2016

[Tada 17] Tada, Yuki and Hagiwara, Yoshinobu and Taniguchi, Tadahiro.: Comparative Study of Feature Extraction Methods for Direct Word Discovery with NPB-DAA from Natural Speech Signals, Joint IEEE International Conference on Development and Learning and Epigenetic Robotics, 2017

[Ozaki 18] Ryo Ozaki and Tadahiro Taniguchi.: Accelerated Nonparametric Bayesian Double Articulation Analyzer for Unsupervised Word Discovery, The 8th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics, 2018

[Johnson 13] Johnson, Matthew J. and Willsky, Alan S.: Bayesian Nonparametric Hidden Semi-Markov Models, Journal of Machine Learning Research, Vol.14, pp.673701, Feb.2013