Double Articulation Analyzer with Prosody for Unsupervised Word and Phone Discovery

Yasuaki Okuda, Ryo Ozaki, Soichiro Komura, and Tadahiro Taniguchi

***Abstract*—Word and phone discovery are important tasks in the language development of human infants. Infants acquire words and phones from unsegmented speech signals using segmentation cues such as distributional, prosodic, and co-occurrence information. Many pre-existing computational models designed to represent this process tend to focus on distributional or prosodic cues. In this study, we propose a nonparametric Bayesian probabilistic generative model called the prosodic hierarchical Dirichlet process-hidden language model (prosodic HDP-HLM) designed to perform simultaneous phone and word discovery from continuous speech signals encoded as time-series data that may exhibit a double articulation structure. Prosodic HDP-HLM, as an extension of HDP-HLM, considers both prosodic and distributional cues within a single integrative generative model. We further propose a prosodic double articulation analyzer (Prosodic DAA) based on an inference procedure derived for prosodic HDP-HLM. We conducted three experiments on different types of datasets, including, Japanese vowel sequence, utterances for teaching object names and features, and utterances following Zipf's law, and the results demonstrated the validity of the proposed method. The results show that the Prosodic DAA successfully used prosodic cues and was able to discover words directly from continuous human speech using distributional and prosodic information in an unsupervised manner, outperforming a method that solely used distributional cues. In contrast, the phone discovery performance did not improve. We also show that prosodic cues contributed to word discovery performance more when the word frequency was distributed more naturally, i.e., following Zipf's law.**

***Index Terms*—Bayesian nonparametrics, child development, language acquisition, prosody, word discovery, phone acquisition, Zipf's law**

I. Introduction

SPEECH signal segmentation problems based on the identification of word and phone boundaries from continuous speech using segmentation cues such as distributional and prosodic cues are essential for human infant language acquisition. This task is simplified if speech signals are always given as a single word. However, many infant-directed spoken utterances consist of multiple words [1], [2]. Nevertheless, human infants can discover words and phonemes from raw continuous speech signals[1]. This word discovery from continuous speech signals is a difficult task because infants do not have access to any information that explicitly identifies the boundaries of words but rather only uses cues contained in continuous speech signals, i.e., they perform what would be considered unsupervised learning in the context of machine learning. Similarly, phone discovery must also be performed using speech signals in an unsupervised manner.

Human infants can exploit numerous cues to discover words from continuous speech signals in the language acquisition process [4]. These cues can be divided into several categories, including 1) distributional, 2) prosodic, 3) co-occurrence, and other cues. The distributional cues represent the statistical relationships between successive elements of speech sound. Prosodic cues rely on acoustic information, such as silent pauses, stressed syllables, rhythmic bias, and suprasegmental features, e.g., pitch [5]–[9]. 3) Co-occurrence cues represent events and objects observed in accordance with the utterance of a word. It has been reported that 8-month-old infants can discover words from fluent speech based solely on distributional cues [10], [11]. It has also been reported that 7-month-old infants can discover words from fluent speech based on distributional rather than prosodic cues [12]. In contrast, it has also been reported that 2-month-old infants can perform word discovery using prosodic information [13]. Prosodic cues have been shown to benefit word segmentation in language acquisition [14], [15]. Based on this, Ludusan et al. extended the unsupervised word segmentation method [16] to use prosodic cues, and showed that prosodic cues were useful in unsupervised word segmentation [17], [18]. Several models have also incorporated co-occurrence cues (e.g., object and place categories) in conjunction with distributional cues to achieve word discovery [19]–[26]. Cross-situational learning and visually grounded acoustic unit discovery have been explored in the fields of speech and computational linguistics [22]–[26]. In robotics, unsupervised word discovery methods have been developed to achieve online lexical acquisition and overcome the out-of-vocabulary problem [19]–[21], [27]–[30]. As a result, considering multiple cues is crucial for the development of an unsupervised phone and word discovery model. In this study, we focus on the development of a computational model that integrates distributional and prosodic cues jointly to perform word discovery.

Several statistical unsupervised simultaneous learning methods have been proposed for acoustic and language models [31]–[34]. Word segmentation and phone categorization are mutually dependent. The unsupervised learning of an acoustic model, i.e., phone (or phoneme) discovery, is a task of clustering feature vectors obtained from continuous speech

Y. Okuda, R. Ozaki and S. Komura are with the Graduate School of Information Science and Engineering, Ritsumeikan University, 1-1-1 Noji Higashi, Kusatsu, Shiga 525-8577, Japan {okuda.yasuaki, ryo.ozaki, komura.soichiro}@em.ci.ritsumei.ac.jp

T. Taniguchi is with College of Information Science and Engineering, Ritsumeikan University, 1-1-1 Noji Higashi, Kusatsu, Shiga 525-8577, Japan taniguchi@em.ci.ritsumei.ac.jp

This work was supported by JSPS KAKENHI Grant Numbers JP16H06569,JP21H04904.

[1]Of note, "phone" and "phoneme" segmentation should be distinguished [3]. In this study, we deal with "phone" segmentation.

signals. Mixture models, such as the Gaussian mixture and hidden Markov models, have been used to categorize feature vectors of phones [35]–[39]. Phone acquisition is a complex categorization task in a feature space because of the overlap in the distribution of the feature vectors of each phone. The actual sound of a phone depends on its context; feedback information from segmented words has been reported as playing an important role in phone acquisition [40]. Therefore, simultaneous word and phone discovery are essential.

Several computational models have been developed to perform unsupervised word and phone discovery simultaneously using distributional cues [33], [34], [41]. Taniguchi et al. proposed a probabilistic generative model (PGM) for simultaneous word and phone discovery, called the hierarchical Dirichlet process-hidden language model (HDP-HLM) and an associated inference algorithm [41]. The HDP-HLM is a PGM for time-series data that potentially has a double articulation structure, that is, hierarchically organized latent words and phones embedded in speech signals. An unsupervised learning method called the nonparametric Bayesian double articulation analyzer (NPB-DAA) was proposed based on HDP-HLM. The NPB-DAA can estimate the double articulation structure and discover words and phones that are simultaneously acquired in acoustic and language models. However, their performance remains limited. The Zero Resource Speech Challenges aim to develop a system that learns an end-to-end spoken dialog system using only information available in the language acquisition process [42]–[44]. Probabilistic computational models that achieved unsupervised direct word discovery from continuous speech signals were proposed as part of this challenge. Kamper et al. proposed an unsupervised word segmentation system designed to segment and cluster speech data into a unit, such as a word [45]. Recently, methods involving representation learning have also been developed [46], [47]. Existing words and phone (or phoneme) discovery approaches are mostly based on dynamic time warping (DTW) [48], representation learning learning [47], [49], or PGMs [41], [50].

In this study, we focus on the PGM-based approach. PGM-based methods can describe cognitive and developmental processes that can determine the latent structure of sensorimotor information and the organization of knowledge without human-annotated label data. The cognitive and learning processes are based only on predictions of sensorimotor information (i.e., observations). This is referred to as predictive coding, and this method has recently been attracting attention as a general principle of human cognitive processes. Several Bayesian approaches for understanding human cognition suggest that internal PGMs support perceptual and cognitive processes [51]–[53]. Predictive coding supported by PGMs is also considered as a design principle for cognitive and developmental robots [54], [55]. Therefore, the development of PGMs for word and phone discovery comprises an important scientific and technological contribution to the field of cognitive and developmental systems.

Based on this background, the present work is focused on the introduction of prosodic cues believed to be useful in word discovery as supplemental cues. We propose a PGM called the prosodic hierarchical Dirichlet process-hidden language model (prosodic HDP-HLM) and an associated inference algorithm. Prosodic HDP-HLM jointly makes use of distributional and prosodic cues and performs simultaneous word and phone discovery. The PGM contains both an acoustic and language models, explicitly. The latent double articulation structure (i.e., sequences of phones and words) of speech signals can be analyzed in an unsupervised manner by adopting prosodic HDP-HLM as a generative model of observation data and inferring latent variables. The unsupervised machine learning method is called prosodic double articulation analyzer (Prosodic DAA), which is an extension of NPB-DAA [41]. An overview of the proposed method is presented in Fig. 1). The Prosodic DAA uses distributional and prosodic cues simultaneously in an explicit manner.

Determining the cases in which prosodic cues contribute particularly to word segmentation from the viewpoint of a statistical model is another important problem. If the distribution of words is sufficiently even, distributional cues may be sufficient to perform word segmentation. However, it is widely known that word distribution follows Zipf's law [56]. Zipf's law is an empirical law found in many social, physical, and other scientific domains, which states that the frequency of a word in a given corpus is inversely proportional to its rank in a frequency table of words used in that context. This means that many words are far less frequently observed than other frequently observed words. Naturally, capturing distributional cues from such data is more difficult than capturing data in which every word is observed with equal frequency. We hypothesized that prosodic cues contribute to word discovery performance more when words are distributed more naturally than in artificially prepared datasets, i.e., in adherence to Zipf's law.

The primary contributions of this paper are summarized as follows. 1) We developed a PGM for time-series data including prosody that may exhibit a double articulation structure, and further propose a Prosodic DAA based on a derivation of the inference procedure for prosodic HDP-HLM. 2) We show that the proposed Prosodic DAA can discover words directly from continuous human speech signals using statistical information and prosodic information in an unsupervised manner. 3) We show that prosodic cues contribute to word segmentation more when words are naturally distributed, that is, when their distribution follows Zipf's law.

The remainder of the paper is structured as follows. Section II presents the prosodic HDP-HLM by extending HDP-HLM, describes the inference procedure of prosodic HDP-HLM, and proposes Prosodic DAA. Section III, Section IV, and Section V evaluates the performance of the proposed method using Japanese vowel sequence utterances for teaching object names, features, and utterances following Zipf's law. Section VI concludes the paper.

II. Prosodic DAA

A. Generative Model: Prosodic HDP-HLM

This section describes a PGM, the prosodic HDP-HLM, based on adding auxiliary observations corresponding to prosody to HDP-HLM, which is a nonparametric Bayesian

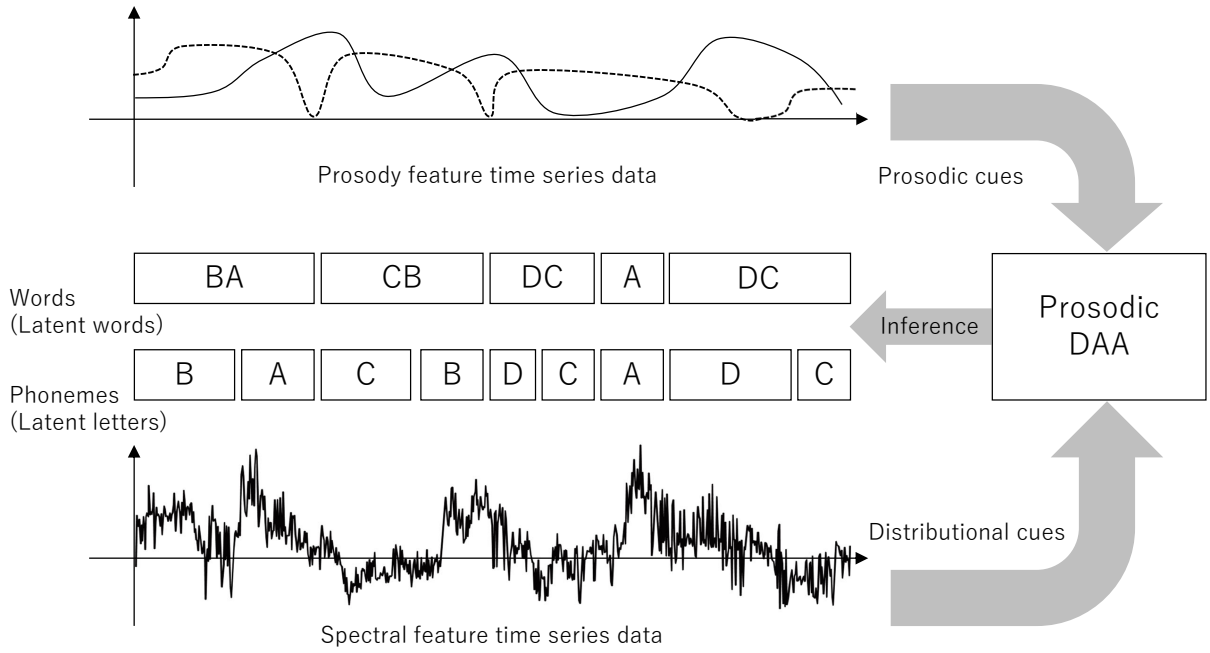


Fig. 1. An illustrative overview of the proposed method

PGM for time-series data that may have a double articulation structure.

A graphical model of the prosodic HDP-HLM is shown in Fig. 2). Notably, most of the generative processes are the same as those of HDP-HLM, except for variables related to prosodic features $\{Y_t\}$.

The generative process of the prosodic HDP-HLM is described as follows.

$$\begin{aligned}
\beta^{\mathrm{LM}} &\sim \mathrm{GEM}(\gamma^{\mathrm{LM}}) && (1)\\
\pi_i^{\mathrm{LM}} &\sim \mathrm{DP}(\alpha^{\mathrm{LM}}, \beta^{\mathrm{LM}}) \quad (\mathrm{i}=1,\ldots) && (2)\\
\beta^{\mathrm{WM}} &\sim \mathrm{GEM}(\gamma^{\mathrm{WM}}) && (3)\\
\pi_i^{\mathrm{WM}} &\sim \mathrm{DP}(\alpha^{\mathrm{WM}}, \beta^{\mathrm{WM}}) \quad (\mathrm{i}=1,\ldots) && (4)\\
w_{i,k} &\sim \pi^{\mathrm{WM}}_{w_{i,k-1}} \quad (i=1,\ldots)(k=1,\ldots,L_i) && (5)\\
(\theta_j, \omega_j) &\sim H \times G \quad (j=1,\ldots) && (6)\\
\phi_q &\sim H_q^{\mathrm{Prosody}} \quad (q=0,1) && (7)
\end{aligned}$$

$$\begin{aligned}
z_s &\sim \pi^{\mathrm{LM}}_{z_{s-1}} \quad (s=1,\ldots,S) && (8)\\
l_{sk} &= w_{w_{z_s k}} \quad (s=1,\ldots,S)(k=1,\ldots,L_{z_s}) && (9)\\
D_{sk} &\sim g(\omega_{l_{sk}}) \quad (s=1,\ldots,S)(k=1,\ldots,L_{z_s}) && (10)\\
x_t &= l_{sk} \quad (t=t^1_{sk},\ldots,t^2_{sk}) && (11)
\end{aligned}$$

$$\begin{aligned}
t^1_{sk} &= \sum_{s'<s} D_{s'} + \sum_{k'<k} D_{sk'} + 1\\
t^2_{sk} &= t^1_{sk} + D_{sk} - 1\\
D^{\mathrm{sum}}_s &= \sum_{s'\le s} D_{s'} && (12)\\
y_t &\sim h(\theta_{x_t}) \quad (t=1,\ldots,T) && (13)
\end{aligned}$$

$$\begin{aligned}
F_t &= \begin{cases} 0 & (t = t^1_{s1} : D^{\mathrm{sum}}_s - 1)\\ 1 & (t = D^{\mathrm{sum}}_s) \end{cases} && (14)\\
Y_t &\sim h^{\mathrm{Prosody}}(\phi_{F_t}) && (15)
\end{aligned}$$

where GEM and DP represent the stick-breaking and Dirichlet processes, respectively. LM represents the language model, and WM represents the word model. The parameters γ^{WM} and α^{WM} are hyperparameters of the word model, β^{WM} is a global transition probability that becomes the base measure of the transition probability distributions, and π_j^{WM} represents the transition probability from latent letter j to the next latent letter. The parameters γ^{LM} and α^{LM} are hyperparameters of the language model, β^{LM} is a global transition probability that becomes the base measure of the transition probability distributions, and π_i^{LM} represents the transition probability from a latent word i to the next latent word. The superscripts LM and WM indicate the language and word models, respectively.

The latent letter[2] sequence of the i-th latent word w_i is sampled from $\pi^{\mathrm{WM}}_{w_{i,k-1}}$. The duration distribution g and observation distribution h have parameters ω_j relating to the j-th latent letter and θ_j generated from the base measures G and H. In addition, the prosodic observation distribution h^{Prosody} has parameters ϕ_q generated from the base measures $H_q^{Prosody}$. The variable z_s is the s-th latent word in the latent word sequence and corresponds to the superstate in the hierarchical Dirichlet process-hidden semi-Markov model (HDP-HSMM) [57]. The duration time D_s is the frame duration of the sth latent word z_s. The latent letter $l_{sk} = w_{z_s k}$ corresponds to the kth latent letter of the sth latent word. The duration time D_{sk} is the frame duration of the latent letter l_{sk}. The duration time D^{sum}_s is the

[2]In the terminology of HDP-HLM, the elements comprising a latent word are called latent letters. In phone and word discovery, a latent letter is considered an inferred index of a phone.

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2022.3210751

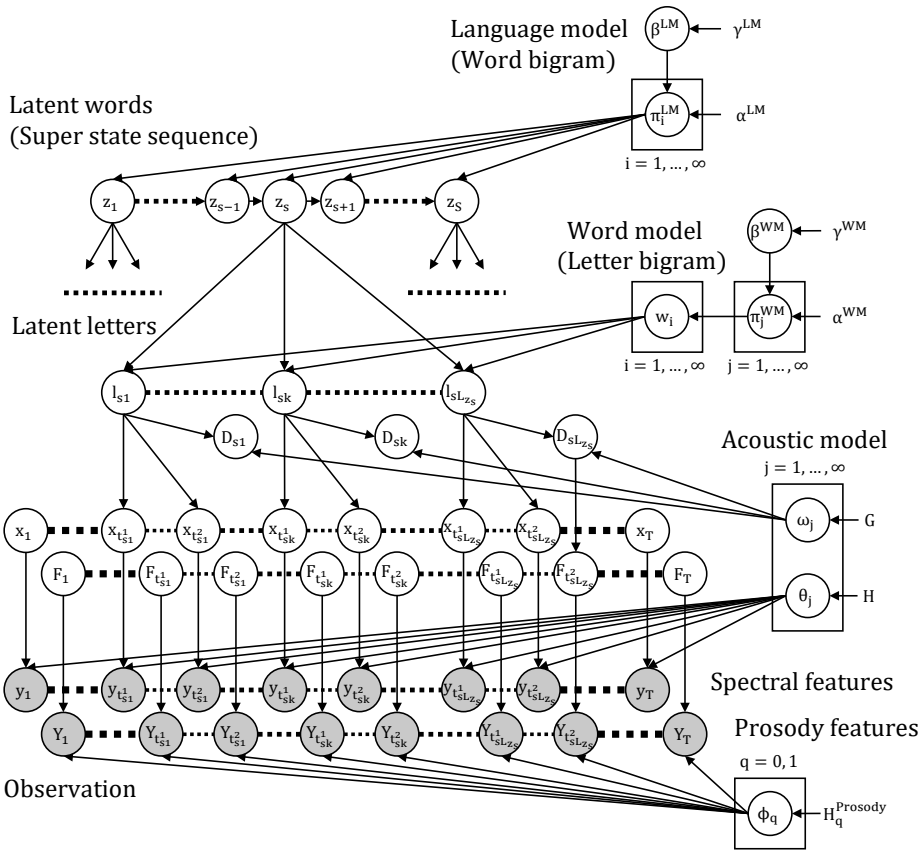


Fig. 2. A graphical model for the proposed prosodic HDP-HLM

frame duration from $t = 1$ to the end point of the word z_s. The variables x_t and y_t indicate the hidden state and observation data at time t, respectively. In word and phone discovery, we assume that y_t represents the spectral feature representation, for example, the mel-frequency cepstral coefficient (MFCC). The time frames t^1_{sk} and t^2_{sk} are the start and end points of the segment corresponding to l_{sk}, respectively.

The duration time D_{sk} of the k-th latent letter l_{sk} of the s-th latent word z_s in the word sequence is drawn from the duration distribution $g(\omega_{l_{sk}})$. The duration of the latent word z_s is $D_s = \sum_{k=1}^{L_{z_s}} D_{sk}$; where g is assumed to be a Poisson distribution, the duration distribution of a latent word z_s also follows a Poisson distribution owing to the reproductive property of this distribution. In this case, the Poisson distribution parameter of the duration of the latent word is $\sum_{k=1}^{L_{z_s}} \omega_{l_{sk}}$.

In addition to the variables described above, which are also in the HDP-HLM, the prosodic HDP-HLM has additional prosody-related variables. The variables Y_t and F_t are prosodic observation data at time t, and indicate that a new word begins at $t+1$ when $F_t = 1$. In this case, $F_t = 1$ when $t = D^{\text{sum}}_s$. The parameter q is a variable relating to the value of 0 or 1 in indicator F_t. We assume that Y_t is a prosodic feature observed in accordance with the word boundaries, i.e., $F_t = 1$.

B. Inference procedure

The approximated blocked Gibbs sampler for the prosodic HDP-HLM can be derived in the same way as the approximated blocked Gibbs sampler for the HDP-HLM. The inference procedure of HDP-HLM, called NPB-DAA, can estimate the double articulation structure from time-series data. The Prosodic HDP-HLM is expected to find latent words and letters from time-series data, including prosody, in an unsupervised manner, by inferring the latent local and global parameters of prosodic HDP-HLM.

In the HDP-HLM, we adopt the backward-filtering forward-sampling procedure, which is the inference method of HDP-HSMM adapted to HDP-HLM. By extending the backward filtering, forward-sampling procedure of HDP-HLM, we obtain an inference procedure for prosodic HDP-HLM. The calculation of the backward messages of the latent word $z_s = i$ in prosodic HDP-HLM is performed as follows.

$$\begin{aligned}
\beta_t(i) &:= p(y_{t+1:T}, Y_{t+1:T} | z_{s(t)} = i, F_t = 1) \\
&= \sum_j \beta^*_t(j) p(z_{s(t+1)} = j | z_t = i) \qquad (16) \\
\beta^*_t(i) &:= p(y_{t+1:T}, Y_{t+1:T} | z_{s(t+1)} = i, F_t = 1) \\
&= \sum_{d=1}^{T-t} \beta_{t+d}(i) p(D_{t+1} = d | z_{s(t+1)} = i) \\
&\quad \times p(y_{t+1:t+d}, Y_{t+1:t+d} | i, d) \qquad (17) \\
\beta_T(i) &:= 1 \qquad (18)
\end{aligned}$$

where $z_{s(t)}$ represents the latent word z_s at time t, and D_{t+1} represents the duration of the latent word beginning at time

Algorithm 1 Inference procedure of Prosodic DAA

Initialize all parameters.
Observe M spectral feature time-series data $\{y^m_{1:T_m}\}_{m\in\{1,2,\ldots,M\}}$ and prosody feature time-series data $\{Y^m_{1:T_m}\}_{m\in\{1,2,\ldots,M\}}$.
repeat
  for $m = 1$ to M **do**
    // Backward-filtering procedure
    For each $i \in \{1,2,\ldots,N\}$, initialize messages $\beta_T(i) = 1$.
    for $t = T$ to 1 **do**
      For each $i \in \{1,2,\ldots,N\}$, compute backward messages $\beta_{t-1}(i)$ and $\beta^*_{t-1}(i)$ using (16)–(18).
    end for
    // Forward-sampling procedure
    Initialize $s = 1$ and $D^{\text{sum}}_s = 0$
    while $D^{\text{sum}_s} < T_m$ **do**
      // Sample a super state representing a latent word using (23).
      $z_s \sim p(z_s \mid y^m_{1:T_m}, Y^m_{1:T_m}, z_{s-1}, F_{D^{\text{sum}}_s} = 1)$
      // Sample duration of the super state using (24).
      $D_s \sim p(D_s | y^m_{1:T_m}, Y^m_{1:T_m}, z_s, F_{D^{\text{sum}}_s} = 1)$
      $D^{\text{sum}}_{s+1} \leftarrow D^{\text{sum}}_s + D_s$
      $s \leftarrow s + 1$
    end while
    $S^m \leftarrow s - 1$
    // Sampling a tentative latent letter sequences
    for $s = 1$ to S^m **do**
      $\bar{w}^m_s \sim P(w | y^m_{D^{\text{sum}}_{s-1}+1:D^{\text{sum}}_s}, \{\pi^{WM}_j, \omega_j, \theta_j\}_{j=1,2,\ldots,J})$
    end for
  end for
  // Update model parameters. See [41].
  Sample acoustic model parameters $\{\omega_j, \theta_j\}$, $\{\phi_q\}$ on the basis of tentatively sampled latent letter sequences $\{\bar{w}^m_s\}$.

  Sample language model parameter $\{\pi^{LM}_i\}, \beta^{LM}$ on the basis of sampled super states , i.e., latent words.
  Sample a word inventory $\{w_i\}_{i=1,2,\ldots,N}$ using SIR procedure.
  Sample a word model $\{\pi^{WM}_i\}, \beta^{WM}$ on the basis of sampled word inventory $\{w_i\}_{i=1,2,\ldots,N}$.
until a predetermined exit condition is satisfied.

$t+1$. The probability $\beta_t(i)$ is obtained by marginalizing all latent words j at time $t+1$. The probability $\beta^*_t(i)$ is the probability that the latent word i begins at time $t+1$. This probability $\beta^*_t(i)$ is obtained by marginalizing all duration frames d. The sum operation in (17) represents marginalization over d. The probability $p(y_{t+1:t+d}, Y_{t+1:t+d}|i,d)$ in (17) shows the probability that observations $y_{t+1:t+d}$ and prosodic observations $Y_{t+1:t+d}$ are generated by the latent word i. The likelihood of the latent word $p(y_{t+1:t+d}, Y_{t+1:t+d}|i,d)$ is as follows.

$$
\begin{aligned}
&p(y_{t+1:t+d}, Y_{t+1:t+d}|i,d) \\
&\quad= \left(\sum_{r\in R^{(L_i,d)}} \prod_{k=1}^{L_i} p(r_k|l_k) \prod_{m=1}^{r_k} p(y_{t+m+\sum_{k'=1}^{k-1} r_{k'}}|l_k) \right) \\
&\qquad\times \left(P(Y_{t+d}|F_{t+d}=1) \prod_{t'=1}^{d-1} P(Y_{t+t'}|F_{t+t'}=0) \right) \qquad (19)
\end{aligned}
$$

$$
R^{(L_i,d)} = \{r \in \{1,2,\ldots\}^{L_i} | \sum_{k=1}^{L_i} r_k = d\} \qquad (20)
$$

where the variable $R^{(L_i,d)}$ is a set of L_i-dimensional natural number vectors whose element summation is d. The value of (19) can be calculated efficiently using dynamic programming. The forward message $\alpha_t(k)$ can be recursively calculated as follows.

$$
\begin{aligned}
\alpha_t(k) &= \sum_{d'=1}^{t-k+1} \alpha_{t-d'}(k-1) p(d'|l_k) \\
&\quad\times \prod_{t'=1}^{d'} p(y_{t-t'+1}, Y_{t-t'+1}|l_k) \qquad (21) \\
\alpha_0(0) &= 1 \qquad (22)
\end{aligned}
$$

where the forward message $\alpha_t(k)$ is defined as the probability that the k-th latent letter in the latent word w_i transitions to the next latent letter at time t. As a result, $\beta_t(i)$ and $\beta^*_t(i)$ can be calculated. The backward-filtering forward-sampling procedure allows the blocked Gibbs sampler to directly sample latent words from observation data without explicitly sampling latent letters in prosodic HDP-HLM, similar to HDP-HLM. In the forward-sampling procedure, the latent word $z_{s(t+1)}$ and duration $D_{s(t+1)}$ of the latent word $z_{s(t+1)}$ are sampled iteratively using backward messages as follows.

$$
\begin{aligned}
&p(z_s = i | y_{1:T}, Y_{1:T}, z_{s-1} = j, F_{D^{\text{sum}}_{1:s}} = 1) = \\
&\qquad p(i|j) \beta_{D^{\text{sum}}_{1:s}}(i) p(y_{D^{\text{sum}}_{1:s}}, Y_{D^{\text{sum}}_{1:s}}|i) \qquad (23)
\end{aligned}
$$

$$
\begin{aligned}
&p(D_s = d | y_{1:T}, Y_{1:T}, z_s = i, F_{D^{\text{sum}}_{1:s}} = 1) = \\
&\quad p(y_{D^{\text{sum}}_{1:s}+1:D^{\text{sum}}_{1:s}+d}, Y_{D^{\text{sum}}_{1:s}+1:D^{\text{sum}}_{1:s}+d} | d, i, F_{D^{\text{sum}}_{1:s}} = 1) \\
&\qquad\times p(d) \frac{\beta_{D^{\text{sum}}_{1:s}+d}(i)}{\beta^*_{D^{\text{sum}}_{1:s}}(i)} \qquad (24)
\end{aligned}
$$

where $D^{\text{sum}}_{1:s} = \sum_{s'<s} D_{s'}$. From the calculation formula shown above, the latent word $z_{s(t+1)}$ and the duration $D_{s(t+1)}$ of the latent word $z_{s(t+1)}$ can be sampled using $\beta_t(i)$ and $\beta^*_t(i)$.

Once the latent words and their duration are sampled, the other parameters (e.g., model parameters and a latent letter sequence for each latent word) can be sampled in exactly the same manner as the original HDP-HLM [41]. For further details, please refer to the original work [41].

C. Prosodic DAA and prosody features

The inference procedure of prosodic HDP-HLM enables the estimation of the structure from time-series data. Therefore, we call the unsupervised machine learning method based on prosodic HDP-HLM as prosodic DAA, in the same way as the unsupervised learning method that is based on the original HDP-HLM is called NPB-DAA[3].

Generally, Prosodic DAA does not specify a feature extraction method for prosody features. Any prosody features that are informative for word boundaries can be used. In the experiment described later in this study, we used the fundamental frequency F_0, and silent pauses were used as prosody observations. We focus on these two prosodic cues because they may be considered as likely universal cues for word discovery [58]. The second-order differential of the fundamental frequency F_0 and the duration of silent pause extracted from audio, instead of being removed, are provided as prosodic feature observations $Y_t := (Y_{1,t}, Y_{2,t})$, respectively. Further details of the feature extraction are described in the section describing the experiments performed.

However, if another prosody feature co-occurring with a word boundary is prepared, such additional prosody features can be easily introduced into Prosodic DAA without any extension of the model.

III. Experiment 1: Continuous Japanese Vowel Speech Signal

In the first experiment, we evaluated Prosodic DAA using Japanese vowel speech signal to verify the applicability of the proposed method to actual human continuous speech signal.

The speech utterances in the dataset exhibit relatively moderate prosody features, that is, they do not have rich prosody features and are monotonous. In addition, the word distributions are artificially designed, and the distributional cues are relatively easy to find. This dataset was used to examine whether the method could find words and phones using distributional cues. In previous studies, NPB-DAA was shown to be able to discover phones and words from this dataset [41], [59]. In this experiment, we evaluated whether the Prosodic DAA could perform word and phone discovery in the same way as NPB-DAA on this dataset. Also, we evaluated whether the Prosodic DAA could improve word and phone discovery even with moderate prosody features.

In this experiment, we compared the proposed method Prosodic DAA with NPB-DAA [41], i.e., statistical word and phone discovery with and without prosodic cues.

A. Conditions

We used the Japanese vowel native speech dataset [4], which was used to evaluate the original NPB-DAA [41], [59]. The data comprised 60 recorded audio files in which a female native Japanese speaker read 30 artificially constructed sentences aloud twice at a natural speed. The sentences comprised five words {aioi, aue, ao, ie, uo}, which consisted of five Japanese vowels {a, i, u, e, o} representing {ä, i, ɯ$^{\beta}$, e̞, o̞} in phonetic symbols respectively. By combining the 5 words, the 30 sentences included 25 two-word sentences, e.g., "aioi," "aue ie," and "uo ao," and five three-word sentences i.e., "aioi uo ie," "aue ao ie," "ao ie ao," "ie uo," and "uo aue ie," were prepared. The set of two-word sentences consisted of all possible word pairs.

The data were encoded into 12-dimensional MFCC time-series data as observation data for spectral features[5]. The frame size of the MFCC was set to 25 ms, and the frame shift of the MFCC was set to 10 ms. We used DSAE as an adaptive feature extractor as described in [59] and extracted 3-dimensional data as observation data and the DSAE parameters $\alpha = 0.003$, $\beta = 0.7$, and $\eta = 0.5$. For more details, please refer to the original paper on NPB-DAA with DSAE [59].

The prosody features were extracted as follows. The second-order differential of the fundamental frequency F_0 ($Y_{1,t}$) and duration of silent pause ($Y_{2,t}$) were extracted and used as time-series data for prosody feature observations. Robust Epoch and Pitch EstimatoR (REAPER[6]) was used to extract the fundamental frequency F_0. The parameter of frame size was set to 0.01. The parameters of minimum and maximum F_0 were set to 40.0 and 300.0, respectively. These parameters were determined empirically. A section where the volume below the threshold is continuous for a certain period was defined as a silent pause. When the duration of a silent pause exceeding a time frame t was detected and extracted, the duration d^{sil} was set to $Y_{1,t} = d^{\mathrm{sil}}$ representing the silent pause between the current frame and the next frame. The threshold of maximum volume and minimum period of silent pause were set to -8 dB and 0.01 s, respectively.

The hyperparameters for HDP-HLM and Prosodic HDP-HLM were set to $\gamma^{\mathrm{LM}} = 10.0$, $\alpha^{\mathrm{LM}} = 10.0$, $\gamma^{\mathrm{WM}} = 10.0$, and $\alpha^{\mathrm{WM}} = 10.0$. The hyperparameters of the duration distribution were set to $\alpha_0 = 200$ and $\beta_0 = 10$. The hyperparameters of the observation distribution were set to $\mu_0 = 0$, $\Sigma_0 = I$, $\kappa_0 = 0{,}01$, and $\nu_0 = (\mathrm{dimension} + 2)$. The hyperparameters of the prosodic observation distribution were set to $\mu_0 = 0$, $\Sigma_0 = I$, $\kappa_0 = 100$, and $\nu_0 = (\mathrm{dimension} + 2)$ for $F_t = 0$, and $\mu_0 = 1$, $\Sigma_0 = I$, $\kappa_0 = 2$, and $\nu_0 = (\mathrm{dimension} + 2)$ for $F_t = 1$. In addition, we set the maximum number of letters and words and the maximum frame duration of words to 10 and 90 for weak-limit approximation. With different random number seeds, 20 trials with a Gibbs sampling of 100 iterations were performed. Regarding the hyperparameters of the HDP-HLM, those used in the original paper were employed. The hyperparameters for the prosodic observations were determined empirically.

B. Results

The phone and word discovery task can be regarded as an unsupervised clustering task. We evaluated the experimental

[3]The source code of Prosodic DAA is available at https://github.com/EmergentSystemLabStudent/Prosodic-DAA

[4]Japanese vowel native speech dataset: https://github.com/EmergentSystemLabStudent/aioi_dataset

[5]In contrast to Experiments 2 and 3, we used 12-dimensional MFCC instead of 36-dimensional MFCC, which contains the first and second derivatives because the derivatives mainly contribute to the recognition of consonants and 12-dimensional MFCC were considered sufficient for Japanese vowel recognition.

[6]REAPER : https://github.com/google/REAPER

TABLE I
ARI AND ABX SCORE FOR ESTIMATED LATENT LETTERS AND WORDS IN EXPERIMENT 1

Method	Phone ARI	Word ARI	Phone ABX score	Word ABX score
1) Prosodic DAA (fundamental frequency F_0 + silent pause)	0.508±0.021	**0.759±0.034**	0.914±0.010	0.936±0.025
2) Prosodic DAA (silent pause)	0.519±0.034	**0.722±0.059**	0.914±0.011	0.955±0.026
3) Prosodic DAA (fundamental frequency F_0)	**0.521±0.029**	0.718±0.096	**0.915±0.010**	**0.958±0.028**
4) NPB-DAA	**0.511±0.030**	0.667±0.093	**0.923±0.008**	**0.963±0.020**

TABLE II
PRECISION, RECALL, AND F-SCORE FOR BOUNDARY, TOKEN, AND TYPE OF ESTIMATED LATENT WORDS IN EXPERIMENT 1.

Method	Boundary Precision	Boundary Recall	Boundary F-score	Token Precision	Token Recall	Token F-score	Type Precision	Type Recall	Type F-score
1)	**0.911±0.047**	**0.989±0.017**	**0.948±0.028**	**0.842±0.083**	**0.893±0.077**	**0.866±0.076**	**0.361±0.091**	**0.98±0.06**	**0.522±0.099**
2)	**0.884±0.072**	**0.976±0.026**	**0.926±0.045**	**0.773±0.135**	**0.813±0.123**	**0.791±0.126**	**0.298±0.073**	0.950±0.087	**0.450±0.091**
3)	0.836±0.068	0.956±0.024	0.891±0.046	0.678±0.141	0.725±0.126	0.699±0.132	0.228±0.045	**0.970±0.071**	0.368±0.063
4)	0.793±0.073	0.962±0.023	0.867±0.048	0.616±0.129	0.683±0.118	0.646±0.121	0.206±0.036	0.940±0.092	0.337±0.052

results using the adjusted rand index (ARI), which quantifies the performance of a clustering task. ARI takes a value of 1 when the clustering result matches the ground truth, and is zero when the data are clustered randomly. We provided phones, i.e., latent letters and word ground truth labels, to all datasets and evaluated the experimental results. Also, the ABX score was calculated as a reference. The ABX discriminability of category x from category y can be defined as the probability that A and X are further apart from B and X when A and X are from category x and B is from category y, according to a dissimilarity function d [42]. We defined the ABX score based on the ABX discriminability. The ABX score evaluates the similarity of acoustic features of segmented spoken utterances classified into a category (i.e., a phone or a word), and of their classification into different categories. The details of the ABX scores used in this study are described in the Appendix. The ABX score evaluates the similarity of acoustic features of segmented spoken utterances classified into a given category (i.e., a phone or a word) are, and the difference of those classified into different categories are. The ABX score was calculated based on the MFCC feature representations (i.e., spectral features). Therefore, the similarity of the prosody information was not considered. Additionally, we calculated the boundary precision, recall, F-score, token precision, recall, and F-score to evaluate the parsing quality, that is, word segmentation. To evaluate the clustering quality, that is, lexicon discovery, type precision, recall, and F-score were also calculated. We followed the definition of the metrics and the calculation procedure used in ZeroSpeech 2020 [7].

In Table I, phone ARI and ABX score (i.e., the average ARI and ABX score for latent letters), word ARI and ABX score (i.e., those for latent words) estimated by NPB-DAA and proposed Prosodic DAA using only silent pause, only F_0 and both F_0 and silent pauses are shown. The ARI for estimated latent letters and words shows how accurately each method estimated latent letters and words, which correspond to phones and words in speech signals. A higher ARI indicates a more accurate estimation of latent variables. The ABX score for estimated latent letters and words shows the extent to which the organized categories differentiate acoustic features belonging to different categories. Note that the ABX score is not based on a comparison between segmentation results and ground truth labels. Therefore, ARI was used the main quantitative evaluation criterion in this experiment. The experimental results are shown as the average score of 20 trials for ARI and ABX.

The results show that the proposed Prosodic DAAs outperformed NPB-DAA at word ARI in all conditions. In contrast, there was almost no difference between Prosodic DAAs in all conditions and NPB-DAA at phone ARI. Comparing Prosodic DAAs in all conditions with NPB-DAA and calculating the t-test at $p = 0.05$, Prosodic DAA using both F_0 and silent pause, and Prosodic DAA using only silent pauses were statistically significantly different from NPB-DAA at word ARI ($p = 3.6 \times 10^{-4}$ and $p = 3.4 \times 10^{-2}$, respectively). There were no statistically significant differences in all combinations of DAAs at phone ARI.

In contrast, NPB-DAA exhibited the highest ABX scores for words and phones, respectively. NPB-DAA does not consider prosody features and categorizes phones and words by considering MFCC features, which are the basis of the ABX score. This suggests that Prosodic DAA discovers better word segments by using prosody information even while diminishing the discriminability of categories from the viewpoint of spectral features.

The statistically significant differences in word ARI between Prosodic DAA using only silent pause and NPB-DAA showed that the word segmentation performance improved when using prosodic cues. Moreover, the statistically significant differences in word ARI between Prosodic DAA using both F_0 and silent pause and Prosodic DAA using only silent pause showed that the word segmentation performance was improved by using both F_0 and silent pause instead of only silent pause.

Table II lists the precision, recall, and F-score for the boundary, token, and type for each method. This result shows that Prosodic DAA consistently outperformed other methods in terms of the segmentation and clustering quality.

[7]ZeroSpeech 2020: https://zerospeech.com/2020/. These metrics are the updated versions of those in ZeroSpeech 2017 [42].

These results show that the Prosodic DAA improved the word discovery performance of NPB-DAA even when the target data had moderate prosody features.

IV. Experiment 2: Japanese Continuous Speech Signal including Prosody

In the second experiment, we evaluated our proposed method using naturally spoken Japanese speech signals containing consonants, ordinal Japanese vocabularies, and richer prosody features than the speech signals used in Experiment 1. This experiment was conducted to verify the applicability of the proposed method to actual human continuous Japanese utterances.

A. Conditions

We prepared the dataset of continuous Japanese utterances[8]. The data consisted of 70 recorded audio files; a male native Japanese speaker read 70 artificial sentences aloud once at a natural speed. The data consisted of sentences that teach names and features of objects, for example, "kore wa omocha," in English "This is a toy" and "yawarakai yo," "It's soft." The sentences comprised 26 words, consisting of 26 Japanese phones. We prepared phones, i.e., latent letters, and word ground truth labels for all datasets and evaluated the relationship between the ground truth labels and estimated latent letters and words using the ARI. We used the automatic annotation tool provided by the Julius GMM to prepare the ground truth labels.

All feature extraction methods and hyperparameters were used in the same way as in Experiment 1, except for the following. The data were encoded as observation data into 36-dimensional MFCC, which is a concatenation of 12-dimensional MFCC, the differential of 12-dimensional MFCC, and second-order differential of 12-dimensional MFCC time-series data. We used DSAE as an adaptive feature extractor in the same manner as [60] and extracted 9-dimensional data as observation data. For further details, please refer to the original work on NPB-DAA with DSAE for natural speech signals [60].

To extract prosody features, the threshold of the maximum volume and minimum period of silent pause was set to -10 dB and 0.03 s, respectively.

Regarding the hyperparameters for HDP-HLM and Prosodic HDP-HLM, the maximum number of letters and words and the maximum frame duration of words were set to 50 and 120 for weak-limit approximation.

In this experiment, we employed an open-source large vocabulary continuous speech recognition engine, Julius[9] [61] as a method for comparison with the proposed Prosodic DAA. The acoustic model of Julius was trained using a large speech dataset in a supervised manner. For the experiment, we used a GMM-based triphone model and a DNN-based triphone model. Note that Julius is not a phone and word discovery system. The results of phone and word segmentation are shown merely for reference.

We prepared two different groups of conditions for Julius. The first group used Julius because it is a generic speech recognition system. In this group, Julius used a generic word dictionary. The second group used the true word dictionary of the dataset. Therefore, in this group, Julius used the true word list and true phone list to perform continuous speech recognition.

B. Results

In Table III, the ARI of latent letters, i.e., phones, and latent words of two different groups of conditions for Julius, NPB-DAA, and proposed Prosodic DAA using only silent pause, only F_0, and both F_0 and silent pauses are shown. The ARI for estimated latent letters and words shows how accurately each method estimated latent letters and words, which correspond to phones and words in speech signals. A higher ARI indicates a more accurate estimation of latent variables. The experimental results of the DAAs showed an average ARI of 20 trials.

The results show that the proposed Prosodic DAAs in all conditions outperformed NPB-DAA at word ARI. In contrast, there was almost no difference between Prosodic DAAs in all conditions and NPB-DAA at phone ARI. Comparing Prosodic DAAs in all conditions with NPB-DAA and calculating the t-test at $p = 0.05$, there were statistically significant differences in all combinations of DAAs at word ARI ($p = 2.4 \times 10^{-13}$, 4.6×10^{-07}, and 3.2×10^{-4} for Prosodic DAA with both features, F_0, and silent pause, respectively). There were no statistically significant differences in any combination of DAAs at phone ARI.

The statistically significant differences in word ARI between Prosodic DAA using either F_0 or silent pause and NPB-DAA showed that the word segmentation performance improved when using prosodic cues. The statistically significant differences in word ARI between Prosodic DAA using both F_0 and silent pause and Prosodic DAA using either F_0 or silent pause showed that the word segmentation performance was improved by using both F_0 and silent pause instead of either F_0 or silent pause. In addition, the statistically significant difference in word ARI between Prosodic DAA using only silent pause and Prosodic DAA using only F_0 showed that the word segmentation performance was improved by using silent pause instead of F_0.

Even though the word segmentation performance was significantly improved compered to NPB-DAA, the ARI for word discovery of Prosodic DAA was still lower than Julius with true word dictionary. However, the performance was competitive with Julius with a generic vocabulary. Phone ARIs of Prosodic DAA was lower than those of Julius. This means Prosodic DAA could find word units with less accurately estimated phone units than Julius, trained with ground-truth labels. Notably, the results of several ARIs of Julius GMM outperformed those of the Julius DNN, likely because the dataset used in Experiment 2 as true labels was annotated using the automatic annotation tools of Julius GMM.

In the same manner as in Experiment 1, NPB-DAA had the second-highest ABX score for word and phone, whereas

[8]Japanese native speech dataset : https://github.com/EmergentSystemLabStudent/object_teaching_dataset

[9]Open-Source Large Vocabulary Continuous Speech Recognition Engine Julius : https://github.com/julius-speech/julius

TABLE III
ARI AND ABX SCORE FOR ESTIMATED LATENT LETTERS AND WORDS IN EXPERIMENT 2

Method	Phone ARI	Word ARI	Phone ABX score	Word ABX score	Trained Acoustic Model	True Word Dictionary
1) Prosodic DAA (fundamental frequency F_0 + silent pause)	**0.369±0.017**	**0.717±0.030**	0.898±0.009	**0.973±0.008**		
2) Prosodic DAA (silent pause)	**0.371±0.017**	**0.644±0.037**	0.894±0.008	0.968±0.007		
3) Prosodic DAA (fundamental frequency F_0)	0.365±0.016	0.607±0.047	**0.906±0.009**	**0.979±0.008**		
4) NPB-DAA	0.365±0.016	0.559±0.050	**0.904±0.007**	**0.973±0.011**		
5) Julius GMM	0.575	0.557	0.783	0.954	✓	
6) Julius DNN	0.474	0.725	0.756	0.952	✓	
7) Julius GMM with true word dictionary	0.677	0.900	0.841	0.965	✓	✓
8) Julius DNN with true word dictionary	0.493	0.825	0.796	0.950	✓	✓

TABLE IV
PRECISION, RECALL, AND F-SCORE FOR BOUNDARY, TOKEN, AND TYPE OF ESTIMATED LATENT WORDS IN EXPERIMENT 2.

Method	Boundary Precision	Boundary Recall	Boundary F-score	Token Precision	Token Recall	Token F-score	Type Precision	Type Recall	Type F-score
1)	**0.703±0.035**	**0.869±0.017**	**0.777±0.024**	**0.332±0.049**	**0.389±0.05**	**0.358±0.049**	**0.263±0.044**	**0.655±0.082**	**0.375±0.057**
2)	**0.681±0.044**	**0.851±0.018**	**0.756±0.028**	**0.270±0.051**	**0.323±0.048**	**0.294±0.049**	**0.220±0.026**	**0.568±0.044**	**0.317±0.033**
3)	0.606±0.033	0.759±0.026	0.673±0.025	0.222±0.056	0.225±0.051	0.223±0.053	0.140±0.032	0.420±0.09	0.210±0.047
4)	0.554±0.048	0.784±0.029	0.648±0.034	0.180±0.046	0.220±0.047	0.197±0.046	0.121±0.025	0.405±0.074	0.186±0.037
5)	0.688	0.932	0.792	0.441	0.542	0.486	0.287	0.833	0.427
6)	0.634	0.876	0.735	0.348	0.402	0.373	0.242	0.733	0.364
7)	0.83	0.98	0.899	0.724	0.749	0.736	0.483	0.967	0.644
8)	0.674	0.871	0.76	0.39	0.425	0.406	0.28	0.7	0.4

Prosodic DAA using F_0 had the highest ABX score. This suggests that Prosodic DAA discovered better word segments by using prosody information even while reducing the discriminability of categories from the viewpoint of spectral features. In particular, the results suggest that the use of silent pauses had larger effect on this phenomenon.

Table IV shows that Prosodic DAA consistently outperformed the other methods in terms of the segmentation and clustering quality.

These results show that the proposed Prosodic DAA may be considered a more effective machine learning approach for estimating the latent double articulation structure of time-series data, including prosody, compared to existing methods.

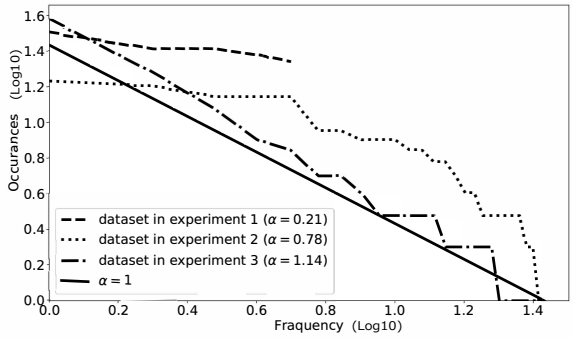


Fig. 3. Log-log plot of the rank—frequency distribution of the datasets used in the experiments. The base of the logarithm for each axis is 10.

V. EXPERIMENT 3: CONTINUOUS JAPANESE SPEECH SIGNALS FOLLOWING ZIPF'S LAW

In the third experiment, we evaluated the Prosodic DAA using continuous, Japanese speech signals more naturalistic than those in Experiment 2 from the viewpoint of distributional properties. It is widely known that words are distributed following Zipf's law, which is a power law, in other words, in documents and utterances [56]. Zipf's law is an empirical law found in many types of social, physical, and other scientific domains. In data satisfying Zipf's law, the rank–frequency distribution has an inverse relation. The frequency of a word is inversely proportional to its rank in the frequency table of words.

$$P(\geq x) \propto x^{-\alpha} \tag{25}$$

where α is a positive constant. In natural language, the word rank–frequency distribution follows $\alpha = 1$ in many cases [62].

The dataset used in Experiments 1 and 2 did not follow Zipf's law. Figure 3 shows the log—log plot of the rank—frequency distributions of datasets for Experiments 1 and 2. This shows that the datasets did not follow Zipf's law. Mathematically, a corpus following Zipf's law has more words that appear less frequently, and the distributional cues for word segmentation are more difficult to capture. Therefore, we hypothesize that prosodic cues contribute significantly to word and phone discovery.

A. Conditions

The dataset used in Experiment 3 was of the same type as that used in Experiment 2, except for its word frequency distribution. The word frequency was adjusted to follow Zipf's

TABLE V
ARI AND ABX SCORE FOR ESTIMATED PHONES AND WORDS IN EXPERIMENT 3

Method	Phone ARI	Word ARI	Phone ABX score	Word ABX score	Trained Acoustic Model	True Word Dictionary
1) Prosodic DAA (fundamental frequency F_0 + silent pause)	0.291±0.016	**0.539±0.039**	0.865±0.009	0.943±0.014		
2) Prosodic DAA (silent pause)	0.292±0.019	**0.478±0.088**	0.869±0.012	**0.947±0.014**		
3) Prosodic DAA (fundamental frequency F_0)	**0.296±0.015**	0.373±0.039	**0.882±0.014**	**0.963±0.019**		
4) NPB-DAA	**0.292±0.015**	0.303±0.040	**0.870±0.012**	0.946±0.017		
5) Julius GMM	0.630	0.822	0.691	0.582	✓	
6) Julius DNN	0.437	0.533	0.633	0.661	✓	
7) Julius GMM with true word dictionary	0.652	0.930	0.672	0.648	✓	✓
8) Julius DNN with true word dictionary	0.471	0.829	0.665	0.652	✓	✓

TABLE VI
PRECISION, RECALL, AND F-SCORE FOR BOUNDARY, TOKEN, AND TYPE OF ESTIMATED LATENT WORDS IN EXPERIMENT 3.

Method	Boundary Precision	Boundary Recall	Boundary F-score	Token Precision	Token Recall	Token F-score	Type Precision	Type Recall	Type F-score
1)	**0.650±0.037**	**0.89±0.064**	**0.75±0.039**	**0.338±0.058**	**0.465±0.095**	**0.391±0.072**	**0.309±0.073**	**0.692±0.124**	**0.427±0.092**
2)	**0.630±0.052**	**0.851±0.101**	**0.721±0.058**	**0.305±0.076**	**0.406±0.133**	**0.346±0.098**	**0.27±0.095**	**0.611±0.181**	**0.373±0.124**
3)	0.553±0.031	0.676±0.028	0.608±0.024	0.166±0.035	0.160±0.034	0.163±0.034	0.120±0.027	0.325±0.07	0.175±0.038
4)	0.503±0.039	0.699±0.038	0.584±0.032	0.122±0.044	0.141±0.045	0.131±0.044	0.096±0.034	0.297±0.089	0.145±0.049
5)	0.711	0.954	0.815	0.577	0.677	0.623	0.338	0.781	0.472
6)	0.605	0.806	0.691	0.303	0.331	0.317	0.227	0.531	0.318
7)	0.803	0.977	0.881	0.691	0.774	0.73	0.508	0.938	0.659
8)	0.627	0.834	0.716	0.338	0.376	0.356	0.279	0.531	0.366

law[10]. Figure 3 shows the log–log plot of the rank–frequency distribution of the dataset used in Experiment 3. The figure shows that the dataset follows Zipf's law, where $\alpha = 1$. A native Japanese male speaker read 42 sentences aloud at natural speed, and the utterances were recorded. The topics of the sentences focused on teaching the names and features of objects in the same way as those in Experiment 2. The sentences consisted of 26 Japanese phones and 27 words.

The other experimental settings were the same as those in Experiment 2.

B. Results

Table V shows ARIs of discovered phones and words. Each shows an average of over 20 trials. The baseline methods are the same as those in Experiment 2.

The experimental results show that the Prosodic DAA improved the word ARI compared with NPB-DAA in every condition significantly at $p = 0.05$ with t-test ($p = 7.2 \times 10^{-21}$, 2.1×10^{-06}, and 1.3×10^{-8} for Prosodic DAA with both features, F_0, and silent pause, respectively). In contrast, significant differences about phone ARI in every condition were not found. ABX scores show tendencies similar to those in Experiments 1 and 2.

Table VI shows that Prosodic DAA consistently outperformed the other methods in terms of the segmentation and clustering quality similarly as in experiments 1 and 2.

This result shows that the prosodic cue contributed to the word segmentation task even when the target data followed Zipf's law. Comparing Tables 2 and 3, we find that the performance of NPB-DAA, which only uses distributional cues, deteriorated. However, we can also find that the introduction of prosodic cues in Experiment 3 improved the ARIs more than those in Experiment 2. This suggests that when prosodic cues contribute to word discovery in natural speech signals, the distributional cues are statistically difficult to capture. In contrast, the performance of Julius with true word dictionary did not deteriorate. In particular, the difference between phone ARIs of DAA-based methods and Julius became larger than that of Experiment 2. This suggests that the distributional property of words can also make phone discovery tasks more difficult.

[10]The Japanese native speech dataset followed Zipf's law : https://github.com/EmergentSystemLabStudent/object_teaching_dataset/

VI. CONCLUSION

In this study, we have proposed a Prosodic DAA designed to discover words directly from continuous human speech signals using statistical and prosodic information in an unsupervised manner. For this purpose, we have proposed a PGM called prosodic HDP-HLM by extending the HDP-HLM. Based on the generative model, we have derived an inference procedure by expanding the blocked Gibbs sampler proposed for HDP-HLM. To evaluate the performance of the proposed method, three experiments were conducted. In the first experiment, we applied the proposed method to actual human Japanese vowel speech signals. In the second experiment, the proposed method was applied to actual human Japanese utterances. In the third experiment, the proposed method was applied to Japanese utterances the word distribution of which followed Zipf's law. The results have shown that the proposed method was able to make use of prosody information and outperformed NPB-DAA in word segmentation performance. However, its performance did not improve over prior methods in terms of phone discovery. This suggests that prosodic cues, i.e.,

second derivatives of F_0 and silent pauses, do not contribute to phone discovery. In addition, the results of the third experiment suggest that prosodic cues contributing to word segmentation of the distributional cues are difficult to capture; for example, whether the word distribution follows Zipf's law.

Word and phone discovery from more natural speech signals remain as a crucial challenge for future research. We have performed word and phone discovery from speech signals. However, we limited the number of words and phones in the experiment. Therefore, we did not test our method on the open-ended learning of words and phones. The TIMIT corpus is widely used to train automatic speech recognition systems, and may be a candidate to which we could apply Prosodic DAA. However, the TIMIT corpus is designed with a small number of repetitions of identical words through unsupervised learning. To achieve simultaneous unsupervised learning of language and acoustic models based on PGMs, the target dataset should have reasonable statistical properties in terms of the distribution of phones and words. Unfortunately, the TIMIT corpus is not suitable for testing the PGM. Thus, applying and adapting the proposed method to a larger dataset remains as a future challenge. In our future work, we intend to focus on the topic of language acquisition from speech signals, emphasizing prosodies such as infant-directed speech by human parents and natural speech signals such as daily conversation.

Computational cost remains as an unsolved problem in scaling up the proposed method to larger datasets. Current inference procedure requires $O(TN^2L_{max}d_{max}^2)$, where T is the number of frames, L_{max} is the maximum number of latent letters in a latent word, d_{max} is the maximum frame length of a latent word, and N_{max} is the maximum number of words [63]. The inference time of the dataset conditioned the maximum number of letters and words to 10 in Experiment 1 and took approximately 4 min, and the dataset conditioned the maximum number of letters and words to 50 in Experiment 2 and took approximately 30 min to perform Gibbs sampling over 100 iterations using two Intel Xeon E5-2650 v2 CPUs at a clock rate of 2.60 GHz, with 8 cores, and 16 threads. The inference time depends on the maximum number of words. Therefore, further improvement in the computational cost is required for a large dataset with many words. Introducing a neural network for inference is expected to be a feasible approach to reduce the computational cost, i.e., amortized inference [64]. This would allow us to make use of GPUs.

The psychological and biological plausibility of the Prosodic DAA will be another topic of discussion. As mentioned in the Introduction, from the viewpoint of predictive coding, PGMs can be regarded as an internal model of the human cognitive system [51]–[54]. Recently, studies have bridged the gap between neuroscience and PGM-based cognitive models to determine how PGMs can be implemented that reflect the processes of the human brain [55], [65], [66].

Prosodic DAA does not specify the method of extracting features of prosody information. However, the types of representations of prosody that can contribute to learning models have been investigated. Recently, representation learning of prosody has also been explored (e.g., [67]). Developing a method that directly uses prosody information (i.e., F0) remains as a future challenge.

In our future research, we intend to aim to develop a robot designed to automatically learn words and phones through human-robot speech interactions. In this study, we focused on language acquisition from statistical information and prosodic information, and proposed a mathematical model of language acquisition. Several studies have suggested that using co-occurrence information improves the accuracy of language acquisition [68], [69]. Moreover, we also intend to consider the combination of co-occurrence cues into Prosodic DAA to obtain a mathematical model for more accurate word and phone discovery.

REFERENCES

[1] R. N. Aslin, J. Z. Woodward, N. P. LaMendola, and T. G. Bever, "Models of word segmentation in fluent maternal speech to infants," in *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, J. L. Morgan and K. Demuth, Eds. Psychology Press, 1995, pp. 117–134.

[2] E. D. Thiessen, E. A. Hill, and J. R. Saffran, "Infant-directed speech facilitates word segmentation," *Infancy*, vol. 7, no. 1, pp. 53–71, 2005.

[3] R. K. Moore and L. Skidmore, "On the use/misuse of the term'phoneme'," *arXiv preprint arXiv:1907.11640*, 2019.

[4] B. Pelucchi, J. F. Hay, and J. R. Saffran, "Statistical learning in a natural language by 8-month-old infants," *Child development*, vol. 80, no. 3, pp. 674–685, 2009.

[5] P. W. Jusczyk, A. Cutler, and N. J. Redanz, "Infants' preference for the predominant stress patterns of english words," *Child development*, vol. 64, no. 3, pp. 675–687, 1993.

[6] P. W. Jusczyk, D. M. Houston, and M. Newsome, "The beginnings of word segmentation in english-learning infants," *Cognitive psychology*, vol. 39, no. 3-4, pp. 159–207, 1999.

[7] J. L. Morgan, "A rhythmic bias in preverbal speech segmentation," *Journal of Memory and Language*, vol. 35, no. 5, pp. 666–688, 1996.

[8] Y. Choi and R. Mazuka, "Young children's use of prosody in sentence parsing," *Journal of psycholinguistic research*, vol. 32, no. 2, pp. 197–217, 2003.

[9] R. Mugitani, T. Kobayashi, A. Hayashi, and L. Fais, "The use of pitch accent in word–object association by monolingual japanese infants," *Infancy*, vol. 24, no. 3, pp. 318–337, 2019.

[10] J. R. Saffran, R. N. Aslin, and E. L. Newport, "Statistical learning by 8-month-old infants," *Science*, vol. 274, no. 5294, pp. 1926–1928, 1996.

[11] E. K. Johnson and P. W. Jusczyk, "Word segmentation by 8-month-olds: When speech cues count more than statistics," *Journal of memory and language*, vol. 44, no. 4, pp. 548–567, 2001.

[12] E. D. Thiessen and J. R. Saffran, "When cues collide: use of stress and statistical cues to word boundaries by 7-to 9-month-old infants." *Developmental psychology*, vol. 39, no. 4, p. 706, 2003.

[13] D. R. Mandel, P. W. Jusczyk, and D. G. K. Nelson, "Does sentential prosody help infants organize and remember speech information?" *Cognition*, vol. 53, no. 2, pp. 155–180, 1994.

[14] A. Christophe and E. Dupoux, "Bootstrapping lexical acquisition: the role of prosodic structure," *The Linguistic Review*, vol. 13, no. 3-4, pp. 383–412, 1996.

[15] J. R. Saffran, E. L. Newport, and R. N. Aslin, "Word segmentation: The role of distributional cues," *Journal of memory and language*, vol. 35, no. 4, pp. 606–621, 1996.

[16] S. Goldwater, T. L. Griffiths, and M. Johnson, "A bayesian framework for word segmentation: Exploring the effects of context," *Cognition*, vol. 112, no. 1, pp. 21–54, 2009.

[17] B. Ludusan, G. Synnaeve, and E. Dupoux, "Prosodic boundary information helps unsupervised word segmentation," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 953–963.

[18] B. Ludusan and E. Dupoux, "The role of prosodic boundaries in word discovery: Evidence from a computational model," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. EL1–EL6, 2016.

[19] T. Nakamura, T. Nagai, K. Funakoshi, S. Nagasaka, T. Taniguchi, and N. Iwahashi, "Mutual learning of an object concept and language model based on MLDA and NPYLM," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2014.

[20] A. Taniguchi, T. Taniguchi, and T. Inamura, "Unsupervised spatial lexical acquisition by updating a language model with place clues," *Robotics and Autonomous Systems*, vol. 99, pp. 166–180, 2018.

[21] A. Taniguchi, Y. Hagiwara, T. Taniguchi, and T. Inamura, "Improved and scalable online learning of spatial concepts and language models with mapping," *Autonomous Robots*, vol. 44, no. 6, pp. 927–946, 2020.

[22] G. Chrupała, L. Gelderloos, and A. Alishahi, "Representations of language in a model of visually grounded speech signal," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 613–622.

[23] H. Kamper, A. Anastassiou, and K. Livescu, "Semantic query-by-example speech search using visual grounding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7120–7124.

[24] D. Harwath, W.-N. Hsu, and J. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," in *The International Conference on Learning Representations (ICLR)*, 2020.

[25] K. Olaleye, B. van Niekerk, and H. Kamper, "Towards localisation of keywords in speech using weak supervision," *arXiv preprint arXiv:2012.07396*, 2020.

[26] D. Harwath, A. Torralba, and J. R. Glass, "Unsupervised learning of spoken language with visual context," in *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, 2016, p. 1866–1874.

[27] T. Araki, T. Nakamura, T. Nagai, S. Nagasaka, T. Taniguchi, and N. Iwahashi, "Online learning of concepts and words using multimodal LDA and hierarchical Pitman-Yor Language Model," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012, pp. 1623–1630.

[28] R. Takeda, K. Komatani, and A. I. Rudnicky, "Word segmentation from phoneme sequences based on Pitman-Yor semi-Markov model exploiting subword information," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 763–770.

[29] R. Takeda and K. Komatani, "Unsupervised segmentation of phoneme sequences based on Pitman-Yor semi-Markov model using phoneme length context," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017, pp. 243–252.

[30] A. Taniguchi, T. Taniguchi, and T. Inamura, "Spatial concept acquisition for a mobile robot that integrates self-localization and unsupervised word discovery from spoken sentences," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 4, pp. 285–297, 2016.

[31] H. Brandl, B. Wrede, F. Joublin, and C. Goerick, "A self-referential childlike model to acquire phones, syllables and words from acoustic speech," in *7th IEEE International Conference on Development and Learning*, 2008, pp. 31–36.

[32] O. Walter, T. Korthals, R. Haeb-Umbach, and B. Raj, "A hierarchical system for word discovery exploiting dtw-based initialization," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 386–391.

[33] C.-y. Lee, T. J. O'donnell, and J. Glass, "Unsupervised lexicon discovery from acoustic input," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 389–403, 2015.

[34] H. Kamper, "Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6535–3539.

[35] G. K. Vallabha, J. L. McClelland, F. Pons, J. F. Werker, and S. Amano, "Unsupervised learning of vowel categories from infant-directed speech," *Proceedings of the National Academy of Sciences*, vol. 104, no. 33, pp. 13 273–13 278, 2007.

[36] B. M. Lake, G. K. Vallabha, and J. L. McClelland, "Modeling unsupervised perceptual category learning," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 1, pp. 35–43, 2009.

[37] C.-y. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.

[38] C.-y. Lee, Y. Zhang, and J. Glass, "Joint learning of phonetic units and word pronunciations for asr," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 182–192.

[39] M.-h. Siu, H. Gish, A. Chan, W. Belfield, and S. Lowe, "Unsupervised training of an hmm-based self-organizing unit recognizer with applications to topic classification and keyword discovery," *Computer Speech & Language*, vol. 28, no. 1, pp. 210–223, 2014.

[40] N. H. Feldman, T. L. Griffiths, S. Goldwater, and J. L. Morgan, "A role for the developing lexicon in phonetic category acquisition." *Psychological review*, vol. 120, no. 4, p. 751, 2013.

[41] T. Taniguchi, S. Nagasaka, and R. Nakashima, "Nonparametric Bayesian double articulation analyzer for direct language acquisition from continuous speech signals." *IEEE Trans. Cognitive and Developmental Systems*, vol. 8, no. 3, pp. 171–185, 2016.

[42] E. Dunbar, X. N. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux, "The zero resource speech challenge 2017," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 323–330.

[43] D. Renshaw, H. Kamper, A. Jansen, and S. Goldwater, "A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge," in *16th Annual Conference of the International Speech Communication Association (Interspeech)*, 2015.

[44] E. Dunbar, J. Karadayi, M. Bernard, X.-N. Cao, R. Algayres, L. Ondel, L. Besacier, S. Sakti, and E. Dupoux, "The zero resource speech challenge 2020: Discovering discrete subword and word units," *arXiv preprint arXiv:2010.05967*, 2020.

[45] H. Kamper, A. Jansen, and S. Goldwater, "A segmental framework for fully-unsupervised large-vocabulary speech recognition," *Computer Speech & Language*, vol. 46, pp. 154–174, 2017.

[46] B. van Niekerk, L. Nortje, and H. Kamper, "Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge," 2020.

[47] H. Kamper and B. van Niekerk, "Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks."

[48] O. Räsänen and M. A. C. Blandón, "Unsupervised discovery of recurring speech patterns using probabilistic adaptive metrics," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 4871–4875.

[49] F. Kreuk, J. Keshet, and Y. Adi, "Self-supervised contrastive learning for unsupervised phoneme segmentation," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 3700–3704.

[50] B. Yusuf, L. Ondel, L. Burget, J. Černockỳ, and M. Saraçlar, "A hierarchical subspace model for language-attuned acoustic unit discovery," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3710–3714.

[51] A. Clark, "Whatever next? predictive brains, situated agents, and the future of cognitive science," *Behavioral and brain sciences*, vol. 36, no. 3, pp. 181–204, 2013.

[52] K. Friston, "The free-energy principle: a unified brain theory?" *Nature reviews neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[53] J. Hohwy, *The predictive mind*. Oxford University Press, 2013.

[54] A. Ciria, G. Schillaci, G. Pezzulo, V. V. Hafner, and B. Lara, "Predictive processing in cognitive robotics: a review," *Neural Computation*, vol. 33, no. 5, pp. 1402–1432, 2021.

[55] T. Taniguchi, H. Yamakawa, T. Nagai, K. Doya, M. Sakagami, M. Suzuki, T. Nakamura, and A. Taniguchi, "A whole brain probabilistic generative model: Toward realizing cognitive architectures for developmental robots," *Neural Networks*, vol. 150, pp. 293–312, 2022.

[56] G. K. Zipf, *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.

[57] M. J. Johnson and A. S. Willsky, "Bayesian nonparametric hidden semi-Markov models," *Journal of Machine Learning Research*, vol. 14, no. Feb, pp. 673–701, 2013.

[58] J. Vaissière, "Perception of intonation," *The handbook of speech perception*, pp. 236–263, 2008.

[59] T. Taniguchi, R. Nakashima, H. Liu, and S. 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.

[60] Y. Tada, Y. Hagiwara, and T. Taniguchi, "Comparative study of feature extraction methods for direct word discovery with npb-daa from natural speech signals," in *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-Epirob)*, 2017.

[61] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," 2001.

[62] Y. Sano, H. Takayasu, and M. Takayasu, "Zipf's law and heap's law can predict the size of potential words," *Progress of Theoretical Physics Supplement*, vol. 194, pp. 202–209, 2012.

[63] R. Ozaki and T. Taniguchi, "Accelerated nonparametric bayesian double articulation analyzer for unsupervised word discovery," in *The 8th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics 2018*, 2018, pp. 238–244.

[64] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[65] A. Taniguchi, A. Fukawa, and H. Yamakawa, "Hippocampal formation-inspired probabilistic generative model," *Neural Networks*, vol. 151, pp. 317–335, 2022.
[66] A. Taniguchi, M. Muro, H. Yamakawa, and T. Taniguchi, "Brain-inspired probabilistic generative model for double articulation analysis of spoken language," in *IEEE International Conference on Development and Learning (ICDL)*, 2022, pp. 107–114.
[67] Z. Hodari, C. Lai, and S. King, "Perception of prosodic variation for speech synthesis using an unsupervised discrete representation of F0," in *The 10th International Conference on Speech Prosody*, 2020.
[68] A. Taniguchi, T. Taniguchi, and T. Inamura, "Unsupervised spatial lexical acquisition by updating a language model with place clues," *Robotics and Autonomous Systems*, vol. 99, pp. 166–180, 2018.
[69] T. Nakamura and T. Nagai, "Ensemble-of-concept models for unsupervised formation of multiple categories," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 4, pp. 1043–1057, Dec 2018.

APPENDIX

The ABX score used in this study is defined following [42] with some modification. The idea is based on the ABX task, which is inspired by match-to-sample tasks used in human psychophysics. The ABX task is a simple method to measure the discriminability between two sound categories (e.g., phones or words). The ABX discriminability of category x from category y is the probability that A and X are further apart from B and X according to some distance d over the acoustic features for these sounds when A and X are from category x, and B is from category y. Provided that a set of sounds $S(x)$ from category x and a set of sounds $S(y)$ from category y, the probability $\hat{\theta}(x,y)$ is defined as follows.

$$\hat{\theta}(x,y) = \frac{1}{m(m-1)n} \sum_{\substack{a\in S(x)\\ b\in S(y)\\ x'\in S(x)\setminus\{x'\}}} \left(\mathbb{1}_{d(a,x)<d(b,x)} + \frac{1}{2}\mathbb{1}_{d(a,x)=d(b,x)}\right) \tag{26}$$

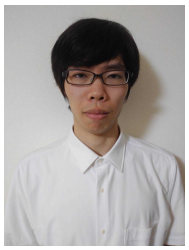
where m and n are the number of sounds in $S(x)$ and $S(y)$, respectively. The function $\mathbb{1}$ denotes an indicator function. The cosine distance is used for d.

The compound measure, i.e., ABX score, is simply averaged over all pairs of categories (i.e., phones and words). If $\#(S(x)) \leq 1$, the category is ignored in the calculation of the ABX score because θ cannot be calculated. Note that the ABX score is calculated for all possible pairs of sound categories[11].

[11]In several studies, ABX scores were calculated on predefined minimal pairs. In such cases, the measure depends on the selected pairs. To avoid arbitrariness caused by the selection, we calculated the ABX score for all possible pairs in this study.

Yasuaki Okuda received a BE degree in Information Science and Engineering from Ritsumeikan University in 2019, and an ME degree from the Graduate School of Information Science and Engineering from Ritsumeikan University in 2021. His current research interests include machine learning and language acquisition.

Ryo Ozaki received a BE degree in Information Science and Engineering from Ritsumeikan University in 2018, and an ME degree from the Graduate School of Information Science and Engineering from Ritsumeikan University in 2020. His current research interests include machine learning and language acquisition.



Soichiro Komura received a BE degree in Information Science and Engineering from Ritsumeikan University in 2021. His current research interests include machine learning and language acquisition.

Tadahiro Taniguchi received his M.E. and Ph.D. degrees from Kyoto University in 2003 and 2006, respectively. He was a Japan Society for the Promotion of Science Research Fellow at the same university from 2005 to 2008. He was an Assistant Professor at the Department of Human and Computer Intelligence, Ritsumeikan University from 2008 to 2010. He was an in the same department from 2010 to 2017. He was a Visiting Associate Professor at the Department of Electrical and Electronic Engineering, Imperial College London from 2015 to 2016. Since 2017, he has been a Professor at the Department of Information and Engineering, Ritsumeikan University and a Visiting General Chief Scientist at the Technology Division of Panasonic Corporation. He has been engaged in research on machine learning, emergent systems, intelligent vehicles, and symbol emergence in robotics.