2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Multimodal Concept and Word Learning Using Phoneme Sequences with Errors

Tomoaki Nakamura[1], Takaya Araki[1], Takayuki Nagai[1],
Shogo Nagasaka[2], Tadahiro Taniguchi[2] and Naoto Iwahashi[3]

***Abstract*— In this study, we propose a method for concept formation and word acquisition for robots. The proposed method is based on multimodal latent Dirichlet allocation (MLDA) and the nested Pitman-Yor language model (NPYLM). A robot obtains haptic, visual, and auditory information by grasping, observing, and shaking an object. At the same time, a user teaches object features to the robot through speech, which is recognized using only acoustic models and transformed into phoneme sequences. As the robot is supposed to have no language model in advance, the recognized phoneme sequences include many phoneme recognition errors. Moreover, the recognized phoneme sequences with errors are segmented into words in an unsupervised manner; however, not all words are necessarily segmented correctly. The words including these errors have a negative effect on the learning of word meanings. To overcome this problem, we propose a method to improve unsupervised word segmentation and to reduce phoneme recognition errors by using multimodal object concepts. In the proposed method, object concepts are used to enhance the accuracy of word segmentation, reduce phoneme recognition errors, and correct words so as to improve the categorization accuracy. We experimentally demonstrate that the proposed method can improve the accuracy of word segmentation and reduce the phoneme recognition error and that the obtained words enhance the categorization accuracy.**

I. Introduction

A key feature of human intelligence is the ability to categorize things. Humans can retrieve information without referring to all of their experiences by using information categories. Furthermore, categorization enables humans to infer unobservable information; such inference can in turn be used to recognize an environment flexibly. Moreover, categories are considered as concepts, and humans can understand the meaning of a word by connecting the word to a concept. We consider such categorization to be of great importance for realizing intelligent robots.

In this light, we have proposed multimodal latent Dirichlet allocation (MLDA) [1], an extension of latent Dirichlet allocation (LDA) [2], and we have shown that a robot can accurately perform categorization by using multimodal information. Furthermore, we have shown that a robot can understand the meanings of words by connecting words to concepts formed by multimodal categorization [3]. Our proposed model enables a robot to infer unobservable information by using a previously learnt model. For example, the robot can infer the sound or hardness of an object from only its appearance. Moreover, it can recall words to represent sensory information perceived by it.

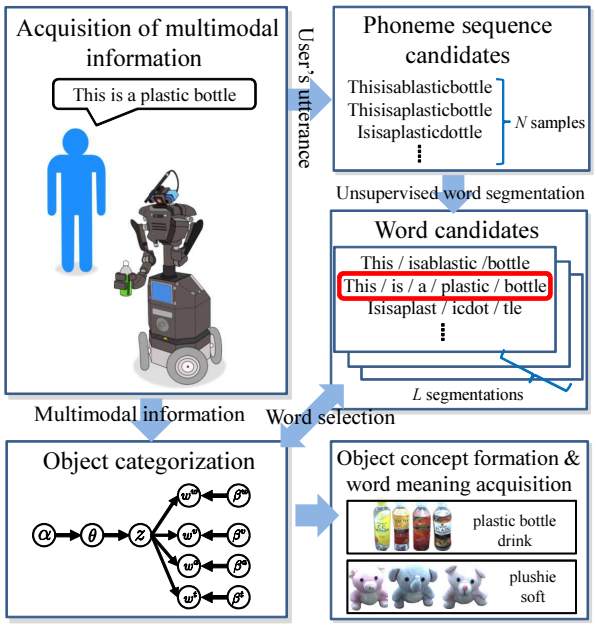


Fig. 1. Overview of proposed method

However, the predefined lexicon used in these methods is one of their disadvantages. We used continuous speech recognition to recognize a user's utterances and a Japanese morphological analyzer to segment sentences into words. Consequently, the proposed method cannot deal with words that are not contained in the lexicon of the speech recognizer and the morphological analyzer. In contrast, humans can build a lexicon by segmenting phoneme sequences according to a phoneme transition probability. We believe that such an ability is also important for robots to acquire language flexibly.

Therefore, we have applied phoneme recognition without a language model and the nested Pitman-Yor language model (NPYLM) [4] to multimodal categorization [5]. A user's utterances are converted into phoneme sequences; then, these are segmented into words using NPYLM in an unsupervised manner; and finally, the segmented words are connected to concepts formed by multimodal categorization. This process enabled the robot to acquire words without using the predefined lexicon. However, other important problems arise. One of them is phoneme recognition errors. It is difficult to correctly recognize the user's utterances by using only phoneme recognition without any language model. The other problem is false segmentation of sentences. NPYLM requires

[1]Dept. of Mechanical Engineering and Intelligent Systems, The University of Electro-Communications, 1-5-1 Chofugaoka Chofu-shi, Tokyo 182-8585, Japan {naka_t, taraki, tnagai}@apple.ee.uec.ac.jp
[2]Dept. of Human and Computer Intelligence, Ritsumeikan University, 1-1-1 Noji Higashi, Kusatsu, Shiga 525-8577, Japan {s.nagasaka, taniguchi}@em.ci.ritsumei.ac.jp
[3]National Institute of Information and Communications Technology, 3-5 Hikari-dai, Seika-cho, Sohraku-gun, Kyoto-fu 619-0289, Japan naoto.iwahashi@nict.go.jp

a large number of sentences to be able to segment the sentences correctly. It is considered difficult to collect a large number of sentences for dialogue between a user and the robot in order to teach object features.

Therefore, in this study, we propose a novel method to enhance the accuracy of phoneme recognition and word segmentation from a limited number of sentences by using MLDA. Here, it is important to note that the words represent the concepts that are formed by a physical robot and that the words and the concepts are interrelated inseparably. Therefore, objects that are classified into the same category are likely to be given the same word with a high probability; furthermore, objects that are connected to the same word are likely to have common features with a high probability. When the robot forms an object concept and segments the user's utterances, which are recognized by phoneme recognition, it can use such clues to improve the accuracy of object categorization, phoneme recognition, and word segmentation mutually.

Unsupervised word segmentation is one of the most important problems in the area of natural language processing (NLP). However, NLP by itself cannot solve semantic segmentation because it requires the grounding of words in the physical world. Our approach has an important perspective in that the physical robot helps to solve the problem of NLP and vice versa.

Figure 1 shows an overview of the proposed method. First, the robot acquires haptic, visual, and audio information by grasping, observing, and shaking an object. Simultaneously, the user teaches the robot object features. Then, the user's utterances are converted into phoneme sequences by phoneme recognition. At this point, multiple candidates of phoneme sequences are obtained by the n-best phoneme recognition results. After that, they are segmented into words. Again, multiple word candidates are obtained using NPYLM with random initial parameters. Therefore, we can obtain $N \times L$ word candidates if the N phoneme sequences are segmented into words L times. Next, the segmented words are connected to multimodal concepts, and the significant words are selected according to a probability that denotes the likelihood of the word representing the concept. Finally, concepts are formed from the multimodal information and the selected words. By using the proposed method, a concept can enhance the word segmentation accuracy and the selected words can enhance the categorization accuracy.

Many studies have focused on categorization using only visual information [6], [7], [8], [9]; however, some categories cannot be formed using only visual information. Some other studies have focused on categorization using the sound made by the robot touching an object [10]; these studies indicate that it is possible to form categories that cannot be formed using only visual information; however, these studies do not consider inference among modalities and word acquisition. We consider that multimodal information is required to form natural categories for humans.

Some studies have already focused on language acquisition [11], [12]. Taguchi et al. proposed a method for learning the correct phoneme sequences by integrating locations and their names [11]. Zuo et al. have proposed a method to correct phoneme recognition errors by restating a word [12]. However, predefined expressions are used to teach words in these studies, and they do not consider word segmentation.

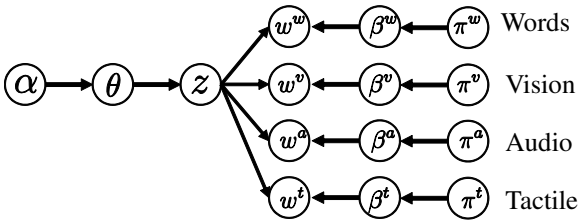


Fig. 2. Graphical model of multimodal LDA

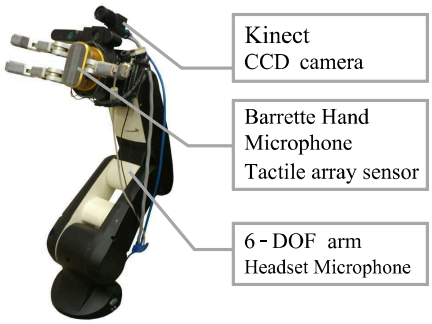


Fig. 3. Robot platform

II. Multimodal Categorization

The robot forms object concepts by using multimodal information captured by grasping, observing, and shaking an object. Figure 2 shows a graphical model of multimodal LDA. In this figure, w^v, w^a, w^t, and w^w respectively denote visual, auditory, haptic, and word information. Details about each type of information are described later. w^v, w^a, w^t, and w^w are respectively generated from a multinomial distribution parameterized by β^v, β^a, β^t, and β^w, which are determined by the Dirichlet prior distribution whose parameters are π^*. Furthermore, θ denotes the parameter of the multinomial distribution that represents the probability of occurrence of category z, and the Dirichlet distribution, whose parameter is α, is assumed to be a hyperparameter of θ. Object categorization is realized by estimating these parameters in the model from objects to be learnt.

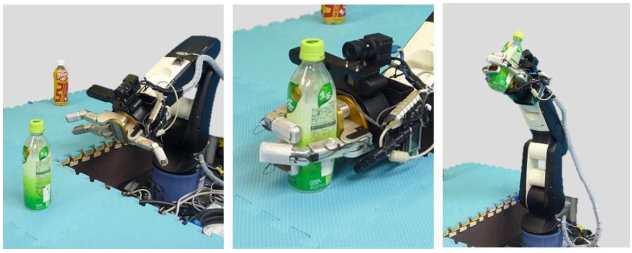
A. Multimodal Information

Multimodal information is captured by the robot shown in Fig.3. In this section, the multimodal information is described in detail.

Visual Information A CCD (charge-coupled device) camera and a depth camera are mounted on the arm of the robot (Fig.3). The robot can observe the object from various viewpoints by moving its arm (Fig.4(a)), and 7 images are captured from $-30°$ to $30°$ in steps of $10°$.

Dense scale-invariant feature transform (DSIFT) [13] is used as a feature calculated from the captured image. Many feature vectors (DSIFT descriptors) are obtained from the image. These feature vectors are vector quantized by using a 500-dimensional codebook. Finally, a 500-dimensional histogram that represents the occurrence frequency of 500 vectors in the codebook is used as visual information.

Haptic Information The robotic hand mounted on the arm and a tactile array sensor mounted on the hand are used to



(a) (b) (c)

Fig. 4. (a) Capture of visual information, (b) Capture of haptic information, and (c) Capture of auditory information

obtain haptic information (Fig.4(b)). The tactile array sensor consists of 162 tactile sensors, and 162 time series sensor values are obtained by the hand grasping the object. The sensor values are approximated by using a sigmoid function, the parameters of which are used as feature vectors for the haptic information [14]. The robot grasps the object five times, and therefore, $810(= 162 \times 5)$ sensor values are obtained from one object. Finally, these feature vectors are vector quantized by using a 15-dimensional codebook, and a 15-dimensional histogram is used as haptic information.

Auditory Information Sound is captured using a microphone mounted on the robot's hand while the robot shakes the object, as shown in Fig.4(c). The sound is divided into frames, and 13-dimensional MFCCs (Mel-frequency cepstrum coefficients) are calculated from each frame. By this process, the frames are transformed into 13-dimensional feature vectors. Finally, as in the case of the visual information, these feature vectors are vector quantized by using a 50-dimensional codebook, and a 50-dimensional histogram is used as auditory information.

Word Information The user teaches the robot object features through speech while the robot is observing the object. Continuous speech signals are converted to phoneme sequences through phoneme recognition, and each sentence is segmented into words in an unsupervised manner by using NPYLM. Finally, a histogram that denotes the occurrence frequency of the words is used as word information.

B. Object Concept Formation

Object categorization is realized by estimating the parameters of the graphical model shown in Fig.2 from the multimodal information that is obtained by the robot. The model parameters are estimated by using Gibbs sampling. A category z_{mij} that is assigned to the i-th feature of modality m of the j-th object is sampled from the posterior probability:

$$p(z_{mij} = k|\boldsymbol{z}^{-mij}, \boldsymbol{w}^m, \alpha, \pi^m) \propto (N_{kj}^{-mij} + \alpha)\frac{N_{mw^mk}^{-mij} + \pi^m}{N_{mk}^{-mij} + W^m\pi^m}, \quad (1)$$

where W^m denotes the dimension of the modality m information and N_{mw^mkj}, the number of times that category k is assigned to w^m of the j-th object. Therefore, N_{mw^mk}, N_{kj}, and N_{mk} are written as follows:

$$N_{mw^mk} = \sum_j N_{mw^mkj}, \quad (2)$$

$$N_{kj} = \sum_{m,w^m} N_{mw^mkj}, \quad (3)$$

$$N_{mk} = \sum_{w^m,j} N_{mw^mkj}. \quad (4)$$

Here, N_{mw^mk} denotes the number of times a category k is assigned to w^m in all objects; N_{kj}, the number of times that category k is assigned to all features in the j-th object; and N_{mk}, the number of times that category k is assigned to the features of modality m information in all objects. The subscript with the minus sign in Eq.(1) indicates the exception of a feature, e.g., $\boldsymbol{z}^{-mij}$ denotes the set of categories that are assigned to all features except for the i-th feature of modality m of the j-th object.

An assignment of a category of the i-th feature of modality m of the j-th object is sampled according to Eq.(1) using Gibbs sampling. By iterating the sampling, the parameters N_* are converged to $\bar{N}_*$. Finally, the estimated parameters $\hat{\beta}_{w^mk}^m$ and $\hat{\theta}_{kj}$ are calculated as follows:

$$\hat{\beta}_{w^mk}^m = \frac{\bar{N}_{mw^mk} + \pi^m}{\bar{N}_{mk} + W^m\pi^m}, \quad (5)$$

$$\hat{\theta}_{kj} = \frac{\bar{N}_{kj} + \alpha}{N_j + K\alpha}. \quad (6)$$

The algorithm for parameter estimation is summarized in Algorithm 1.

Algorithm 1 Algorithm for parameter estimation

```
Iterate the following until convergence
for all m, i, j do
    u ← random value [0, 1]
    for k ← 1 to K do
        P[k] ← P[k − 1] + (N_kj^{−mij} + α) (N_{mw^m k}^{−mij} + π^m) / (N_mk^{−mij} + W^m π^m)
    end for
    for k ← 1 to K do
        if u < P[k]/P[K] then
            z_mij = k, break
        end if
    end for
end for
```

C. Category Recognition of Unseen Object

By using the learnt MLDA model, it is possible to recognize the category of an unseen object. The category is determined by calculating the probability that the unseen object is classified into each category from the multimodal information obtained from it. When obtaining multimodal information $\boldsymbol{w}_{obs}^v$, $\boldsymbol{w}_{obs}^a$, and $\boldsymbol{w}_{obs}^t$, the category z is determined by maximizing the probability $P(z|\boldsymbol{w}_{obs}^v, \boldsymbol{w}_{obs}^a, \boldsymbol{w}_{obs}^t)$. Therefore, the category z of an unseen object is calculated as follows:

$$\hat{z} = \operatorname*{argmax}_z P(z|\boldsymbol{w}_{obs}^v, \boldsymbol{w}_{obs}^a, \boldsymbol{w}_{obs}^t) = \operatorname*{argmax}_z \int P(z|\theta)P(\theta|\boldsymbol{w}_{obs}^v, \boldsymbol{w}_{obs}^a, \boldsymbol{w}_{obs}^t)d\theta \quad (7)$$

In this equation, $P(\theta|\boldsymbol{w}^v_{obs}, \boldsymbol{w}^a_{obs}, \boldsymbol{w}^t_{obs})$ is required to be recalculated depending on the unseen object, and it is obtained by the same algorithm as that used in the learning phase by using the following equation instead of Eq.(1).

$$p(z_{mij} = k|\boldsymbol{z}^{-mij}, \boldsymbol{w}^m, \alpha, \pi^m) \propto (N^{-mij}_{kj} + \alpha)\frac{\bar{N}_{mw^mk} + N^{-mij}_{mw^mk} + \pi^m}{\bar{N}_{mk} + N^{-mij}_{mk} + W^m\pi^m} \quad (8)$$

where $\bar{N}_{mw^mk}$ and $\bar{N}_{mk}$ are converged values in the learning phase.

D. Inference of Unobservable Information

The proposed multimodal categorization enables not only the classification of objects but also the inference of the unobservable information of an object. The robot can infer the properties of the object only from visual information, such as how hard it is, whether it makes a sound, what sound it makes, and which word it is represented by. For example, we consider the case of the inference of word information from only visual information. The probability that word information w^w is generated from only visual information $\boldsymbol{w}^v_{obs}$ is written as follows:

$$p(w^w|\boldsymbol{w}^v_{obs}) = \int \sum_z p(w^w|z)p(z|\theta)p(\theta|\boldsymbol{w}^v_{obs})d\theta. \quad (9)$$

$p(\theta|\boldsymbol{w}^v_{obs})$ can also be recalculated in the same manner as in the previous section.

III. Unsupervised Word Segmentation

The lexicon has been assumed to be predefined, and we used the morphological analyzer that we had previously used in our research [3]. As mentioned earlier, there exists the problem that the robot cannot deal with words that are not contained in the dictionary of the morphological analyzer. To overcome this problem, we use the nested Pitman-Yor language model, using which sentences are segmented into words in an unsupervised manner.

A. Hierarchical Pitman-Yor Language Model

The hierarchical Pitman-Yor language model (HPYLM) is an n-gram model with a hierarchical Pitman-Yor process. The probability that word w appears after context h is written as follows:

$$p(w|h) = \frac{c(w|h) - d \cdot t_{hw}}{\theta + \sum_w c(w|h)} + \frac{\theta + d \cdot \sum_w t_{hw}}{\theta + \sum_w c(w|h)} p(w|h') \quad (10)$$

where h' denotes the $(n-1)$-gram. Therefore, $p(w|h')$ is the probability that w appears after the context, which is one shorter context than h, and it can be computed recursively. Furthermore, $c(w|h)$ denotes the number of occurrences of w, which is generated from context h, and t_{hw}, the number of occurrences of w, which is generated from context h', in $c(w|h)$. The parameters d and θ are estimated by using Gibbs sampling.

B. Nested Pitman-Yor Language Model

If the lexicon is given, the probability $p(w|h')$ can be set to the inverse of the number of words. However, it is difficult to calculate this probability without a predefined dictionary, because all substrings could be a word. NPYLM solves this problem by using a character HPYLM as a base measure of the word unigram. This model is called the nested Pitman-Yor language model (NPYLM) because the character HPYLM is embedded as a base measure of the word HPYLM. NPYLM can rapidly segment sentences by using a blocked Gibbs sampler and dynamic programming.

IV. Word Meaning Acquisition

NPYLM can be used to segment sentences in an unsupervised manner. If a large number of sentences that do not include errors are obtained, NPYLM can segment these sentences into words with high accuracy [4]. However, it is considered difficult to obtain many sentences for learning when a user teaches words that represent object features to the robot. Furthermore, it is difficult to obtain a phoneme sequence without errors. Moreover, another problem exists in that the segmentation result varies each time because the learning of NPYLM is based on the sampling. To overcome these problems, we use the object concept, which is formed by MLDA. First, multiple candidates for words are computed by applying NPYLM to n-best phoneme recognition results many times. Then, words that are strongly connected to the object concept are selected by using multimodal categorization. Therefore, the robot can obtain words that represent the objects and their features. The detailed algorithm for this procedure is as follows.

1) An i-th user's utterance given for a j-th object are recognized by using phoneme recognition, and N-best phoneme sequences $p_{jin}(1 \leq n \leq N)$ are obtained.
2) The phoneme sequences p_{jin} are segmented into words by applying NPYLM L times, and word histograms $\bar{\boldsymbol{w}}^w_{jinl}(1 \leq l \leq L)$, which represent the occurrence frequency of the words, are computed. Therefore, $N \times L$ candidates of a word histogram of the i-th user's utterance given for the j-th object are obtained, and a set of the candidates of word histograms of all objects is denoted by $\bar{\boldsymbol{W}}^w = \{\bar{\boldsymbol{w}}^w_{jinl}|1 \leq j \leq J, 1 \leq i \leq I_j, 1 \leq n \leq N, 1 \leq l \leq L\}$. I_j denotes the number of user's utterances for the j-th object.
3) Iterate the following for all $j(= 0, \cdots, J)$
   i) Object concepts are formed by using the multimodal information $\boldsymbol{W}^v_{-j}, \boldsymbol{W}^a_{-j}, \boldsymbol{W}^t_{-j}$, and $\bar{\boldsymbol{W}}^w_{-j}$, where a subscript with a minus sign indicates the exception of the j-th object's information from $\boldsymbol{W}^v, \boldsymbol{W}^a, \boldsymbol{W}^t$, and $\bar{\boldsymbol{W}}^w$, respectively.
   ii) A histogram $\boldsymbol{w}^w_{ji}$ that represents the occurrence frequency of words of the i-th sentence given for the j-th object is selected from the candidate with the highest probability among the candidates of word histogram $\bar{\boldsymbol{w}}^w_{jinl}(1 \leq n \leq N, 1 \leq l \leq L)$ as follows:

$$\boldsymbol{w}^w_{ji} = \mathrm{argmax}_{n,l}\, p(\bar{\boldsymbol{w}}^w_{jinl}|\boldsymbol{w}^v_j, \boldsymbol{w}^a_j, \boldsymbol{w}^h_j). \quad (11)$$

   In this equation, the probability that words $\bar{\boldsymbol{w}}^w_{jinl}$ are generated from multimodal information $\boldsymbol{w}^v_j, \boldsymbol{w}^a_j, \boldsymbol{w}^h_j$ is computed by using the model learnt in (i) and Eq.(9).

Fig. 5. Objects used in the experiment

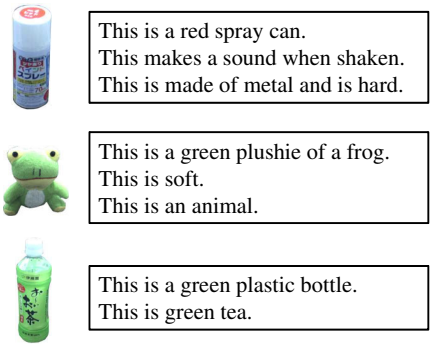


Fig. 6. Examples of the sentences used in the experiment

iii) By summing the word histograms $\boldsymbol{w}_{ji}^{w}$, the word histogram of the j-th object is calculated as follows:

$$\boldsymbol{w}_{j}^{w} = \sum_{i}^{I_j} \boldsymbol{w}_{ji}^{w}. \tag{12}$$

4) Finally, the object concepts are formed by MLDA by using a set of the selected histograms $\boldsymbol{W}^{w} = \{\boldsymbol{w}_{1}^{w}, \boldsymbol{w}_{2}^{w}, \cdots, \boldsymbol{w}_{J}^{w}\}$ and the sets of multimodal information $\boldsymbol{W}^{v}$, $\boldsymbol{W}^{a}$, and $\boldsymbol{W}^{t}$.

A user's utterances can be converted into words that represent object features by integrating unsupervised word segmentation and multimodal categorization as described above.

V. Experiments

An experiment was conducted to validate the proposed method. In the experiment, 50 objects that are classified into 10 categories (e.g., plastic bottles, plushies, etc.) were used. Furthermore, a user taught object features to the robot through speech, and their speech was converted into text through phoneme recognition. The user was a native Japanese speaker, and Japanese sentences were used in this experiment. Fig. 6 shows examples of the sentences used.

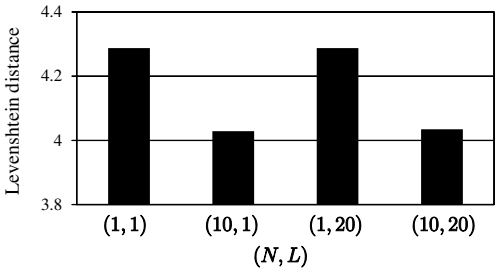


Fig. 7. Average of Levenshtein distance between the selected phoneme sequences and the correct phoneme sequences

TABLE I
THE NUMBER OF WORDS IN THE SELECTED PHONEME SEQUENCES

(N, L)	$(1, 1)$	$(10, 1)$	$(1, 20)$	$(10, 20)$
# of words	255	204	238	208

A. Word Selection

The word selection was conducted by using multimodal information, which were acquired by the robot, and users' utterances. We tested the proposed method by different conditions. The 1-best recognition results ($N = 1$) or 10-best recognition results ($N = 10$) of phoneme recognition were used as candidates of a phoneme sequence. In addition, word candidates were computed from each candidate of the phoneme sequence by applying NPYLM 1 time ($L = 1$) or 20 times ($L = 20$). Therefore, we compared the proposed method under the following four conditions: $(N, L) = (1, 1), (1, 20), (10, 1), (10, 20)$. $(N, L) = (1, 1)$ is the method proposed in [5].

First, we evaluated the difference between the selected phoneme sequences and the correct phoneme sequences. The Levenshtein distance, which represents a string metric for measuring the difference between two sequences, was used as the difference. It is defined as the minimum number of edits needed to transform one string into the other, with the allowable edit operations being insertion, deletion, or substitution of a single character. Fig. 7 shows the average distances between the selected phoneme sequences and the correct phoneme sequences of all user's utterances under each condition. From this figure, the distance of phoneme sequences selected from the candidates of $N = 10$ was shorter than that of $N = 1$. This result indicates that the proposed method can select a more correct phoneme sequence from 10-best phoneme sequences. Tbl. I shows the number of words selected by the proposed method. In the case of $(N, L) = (1, 1)$, which includes only one candidate, the number of words was the largest. This was because the same words were treated as different words owing to false recognition and false segmentation. For example, if "plushie" was recognized as "flushie," this word was treated as a different word from "plushie." On the other hand, in the case of $N = 10$, the number of words was smaller. This was because the likelihood became larger if words that had same meaning were represented by the same phoneme sequence. If the candidates of "flushie" included "plushie", the proposed method was able to select the correct word "plushie."

From these results, the selection of phoneme sequences worked well by integrating multimodal categorization, phoneme recognition, and NPYLM.

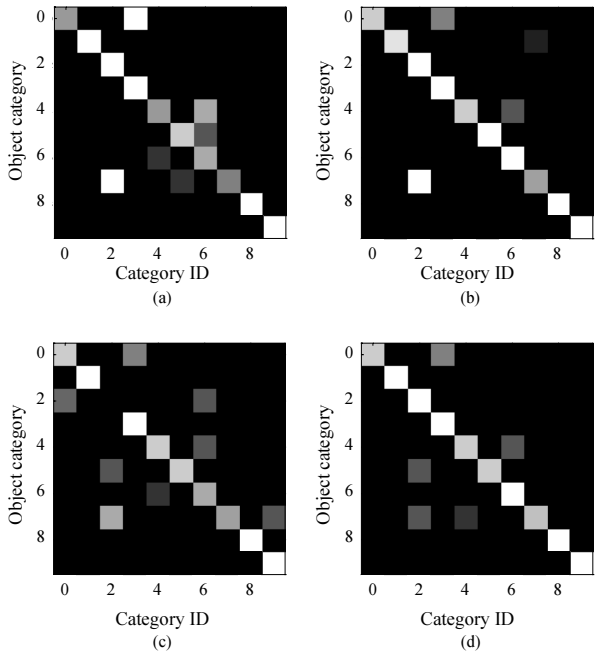


Fig. 8. Classification results by using a word histogram selected from (a) $(N, L) = (1, 1)$, (b) $(N, L) = (10, 1)$, (c) $(N, L) = (1, 20)$, and (d) $(N, L) = (10, 20)$

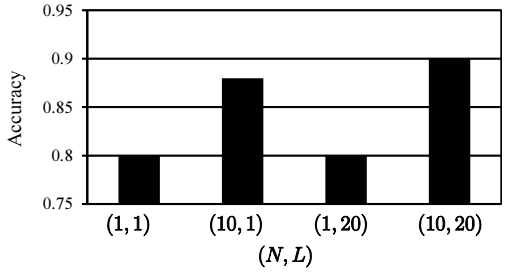


Fig. 9. Classification accuracy by using a word histogram selected from each condition

B. Multimodal Categorization

Multimodal object concepts were formed by using the multimodal information and the words selected in Sec. V-A. Fig. 8 shows the result. In this figure, the vertical axis represents an index of correct categories, the horizontal axis represents an index of a category into which the object is classified, and the brightness reflects the number of objects classified into the category. Furthermore, the classification accuracy under each condition is shown in Fig. 9. We can see that the accuracy under the condition of using 1-best phoneme recognition ($N = 1$) was smaller than that under other cases. It was considered difficult to obtain correct words even though multiple candidates of words were computed by NPYLM ($(N, L) = (1, 20)$) because the word candidates included phoneme recognition errors. On the other hand, in the case of selecting from the 10-best results of phoneme recognition ($(N, L) = (10, 1)$), the accuracy of classification was 88%, because more correct words were selected by the proposed method. Moreover, in the case of computing multiple candidates by applying NPYLM to the 10-best results of phoneme recognition ($(N, L) = (10, 20)$), the accuracy was 90%, which was the largest accuracy in all conditions.

These results indicate that more correct words can be selected by using the multimodal object concepts, and conversely, the objects can be classified more correctly by using the selected words.

VI. Conclusion

In this study, we applied NPYLM to multimodal categorization. Sentences uttered by a user are recognized by phoneme recognition and segmented into words in an unsupervised manner. However, it is difficult to acquire correct words from a limited number of sentences, some of which also include errors. To overcome this problem, many words candidates are computed, and words that represent the object features were selected by the object concept by considering the probabilities that words are generated from the objects.

In future work, we will apply this method to online multimodal categorization [14], which we have studied previously, and we will construct a system that can learn objects more interactively. Moreover, we are considering the use of the learnt language model for speech recognition and a method for improving acoustic models through the learning process.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 23-10330.

References

[1] T. Nakamura, T. Nagai, and N. Iwahashi, "Multimodal Object Categorization by a Robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007, pp. 2415–2420.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] T. Nakamura, T. Nagai, and N. Iwahashi, "Grounding of word meanings in multimodal concepts using LDA," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 3943–3948.

[4] D. Mochihashi, T. Yamada, and N. Ueda, "Bayesian unsupervised word segmentation with nested pitman-yor language modeling," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, vol. 1, 2009, pp. 100–108.

[5] 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*, 2012, pp. 1623–1630.

[6] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering object categories in image collections," in *IEEE International Conference on Computer Vision*, 2005, pp. 17–20.

[7] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, June 2003, pp. 264–271.

[8] L. Fei-Fei, "A bayesian hierarchical model for learning natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 524–531.

[9] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, 2009, pp. 1903–1910.

[10] J. Sinapov and A. Stoytchev, "Object category recognition by a humanoid robot using behavior-grounded relational learning," in *IEEE International Conference on Robotics and Automation*, 2011, pp. 184–190.

[11] R. Taguchi, N. Iwahashi, K. Funakoshi, M. Nakano, T. Nose, and T. Nitta, "Learning lexicons from spoken utterances based on statistical model selection," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 25, no. 4, pp. 549–559, 2010.

[12] X. Zuo, T. Sumii, N. Iwahashi, K. Funakoshi, M. Nakano, and N. Oka, "Correction of phoneme recognition errors in word learning through speech interaction," in *Spoken Language Technology Workshop*, 2010, pp. 360–365.

[13] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *ACM International Conference on Multimedia*, 2010, pp. 1469–1472.

[14] T. Araki, T. Nakamura, T. Nagai, K. Funakoshi, M. Nakano, and N. Iwahashi, "Autonomous acquisition of multimodal information for online object concept formation by a robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 1540–1547.