More-Natural Mimetic Words Generation for Fine-Grained Gait Description

Hirotaka Kato[1(✉)], Takatsugu Hirayama[1], Ichiro Ide[1], Keisuke Doman[2], Yasutomo Kawanishi[1], Daisuke Deguchi[1], and Hiroshi Murase[1]

[1] Nagoya University, Aichi, Japan
katoh@murase.is.i.nagoya-u.ac.jp
[2] Chukyo University, Aichi, Japan

Abstract. A mimetic word is used to verbally express the manner of a phenomenon intuitively. The Japanese language is known to have a greater number of mimetic words in its vocabulary than most other languages. Especially, since human gaits are one of the most commonly represented behavior by mimetic words in the language, we consider that it should be suitable for labels of fine-grained gait recognition. In addition, Japanese mimetic words have a more decomposable structure than these in other languages such as English. So it is said that they have sound-symbolism and their phonemes are strongly related to the impressions of various phenomena. Thanks to this, native Japanese speakers can express their impressions on them briefly and intuitively using various mimetic words. Our previous work proposed a framework to convert the body-parts movements to an arbitrary mimetic word by a regression model. The framework introduced a "phonetic space" based on sound-symbolism, and it enabled fine-grained gait description using the generated mimetic words consisting of an arbitrary combination of phonemes. However, this method did not consider the "naturalness" of the description. Thus, in this paper, we propose an improved mimetic word generation module considering its naturalness, and update the description framework. Here, we define the co-occurrence frequency of phonemes composing a mimetic word as the naturalness. To investigate the co-occurrence frequency, we collected many mimetic words through a subjective experiment. As a result of evaluation experiments, we confirmed that the proposed module could describe gaits with more natural mimetic words while maintaining the description accuracy.

1 Introduction

A mimetic word is used to verbally express the manner of a phenomenon intuitively. The Japanese language is known to have a greater number of mimetic words than most other languages. Researchers have focused on Japanese mimetic words representing the texture of an object to understand the mechanism of cross-modal perception and applied it to information systems [1,2,10]. Human motion, especially gait, is a visually dynamical state most commonly represented

W.-H. Cheng et al. (Eds.): MMM 2020, LNCS 11962, pp. 214–225, 2020.
https://doi.org/10.1007/978-3-030-37734-2_18

by mimetic words, but it has not attracted attention from researchers working on the application of mimetic words to information systems. In English, when we wish to properly express the aspect of gaits, we can use lexical verbs such as *stroll*, *stagger*, and so on. Meanwhile, in Japanese, when we wish to describe the slight difference of gaits, we can use mimetic words adverbially. In addition, Japanese mimetic words have a more decomposable structure than these in other languages. So native Japanese speakers can express them briefly using various mimetic words and even modify them impromptu, in order to express their impressions intuitively.

Japanese mimetic words have an interesting property: sound-symbolism, which indicates that there is an association between linguistic sounds and sensory experiences [5]. The phonemes of a mimetic word should be strongly related to the visual sensation when observing a gait so that the mimetic words can describe the difference in the appearances of gaits at a fine resolution [3]. In the Japanese language, there are more than fifty gait-related mimetic words according to a Japanese mimetic word dictionary [7]. For example, *noro-noro* describes "slowly walk without having a vigorous intention to move forward," and *yoro-yoro* describes "walk with an unstable balance." Their difference of only one sound, i.e. /n/ or /y/, can represent a slight difference in gaits. As another example, *suta-suta* describes "walk with light steps without observing around," and *seka-seka* describes "trot as being forced to hurry." As we can see from these examples, the phoneme /s/ seems to express an impression of fast, smooth, and stable motion. Such associations are individual-invariant and linguistic-invariant similar to the famous Bouba/kiki-effect [8].

We have focused on gaits and proposed a computational method to convert the kinetic features to mimetic words inspired by this cross-modal perception [4]. We constructed a phonetic space simulating the sound-symbolism and associated it with a kinetic feature space of gaits by a regression model. It allows us to describe the difference of gait impressions as difference in phonemes, computationally. Thanks to this ability, the proposed framework can assign not only existing mimetic words but also a novel one generated from an arbitrary combination of phonemes to gaits. However, although it can generate a mimetic word which is closer to one's intuitive impression than ordinary mimetic words, it has a risk of generating useless mimetic words because an extremely uncommon combination of phonemes will sound strange. To avoid this problem, in this paper, we propose an improved word generation module considering its "naturalness". More specifically, we introduce a "naturalness penalty" into the most suitable mimetic word generation module.

The previous study had one more problem that no public dataset was available at that time. So we newly constructed a public dataset. The most notable point of the dataset is that it includes various mimetic words described in a free description form. In this paper, to define what characteristics of words are natural, we analyze these annotations and define the co-occurrence frequency of phonemes composing the mimetic words as the naturalness.

The rest of the paper is composed as follows: Related work is introduced in Sect. 2. Section 3 introduces the dataset. Section 4 introduces our proposed framework briefly and describes the new description module. Section 5 reports results of experiments. Finally, the paper is concluded in Sect. 6.

2 Related Work

Most previous researches focusing on human gaits work on authentication or soft biometrics. For example, Sakata et al. proposed an age estimation method from gaits [9]. There are few studies on the fine-grained description which is independent from individuals. As a study of describing dynamic states, Takano et al. proposed a sentence generation method from RGB-D videos [13]. They introduced a "motion primitive" representation which intermediates motions and sentences. Though their approach is similar to ours in that translating motions to primitive representations, their proposed representation consists of just latent variables which are not intuitively interpretable by people, and the correctness of the representation itself can not be evaluated directly. Meanwhile, in our method, the primitive representations are Japanese mimetic words, and the correctness can be evaluated directly by any native Japanese speaker.

With regard to researches of mimetic words, there are some previous works on mimetic words associated with auditory, visual, and tactile modalities in the field of Computer Science. Sundaram et al. proposed a "meaning space" having the semantic word-based similarity metric that can be used to cluster acoustic features extracted from audio clips tagged with English onomatopoeias (mimetic words of sound) [11]. They also constructed a latent perceptual space using audio clips categorized by high-level semantic labels and the mid-level perceptually motivated onomatopoeia labels [12]. Fukusato et al. proposed a method to estimate an onomatopoeia imitating a collision sound, e.g. "Bang", from the physical characteristics of objects [2]. Shimoda et al. demonstrated that Web images searched with different mimetic words can be classified with a deep convolutional neural network [10]. Doizaki et al. proposed a mimetic word quantification system [1] which is based on sound-symbolism and prior subjective evaluations using 26 opposing pairs of tactile adjectives such as "hard – soft". These works target mimetic words imitating sounds or representing visually static states. Meanwhile, as mentioned in Sect. 1, in this paper, we focus on human gaits as visually dynamic states, especially human gaits, and attempt to accurately describe human gaits using mimetic words.

3 Dataset

We newly constructed a public dataset[1]. It includes videos recording human gaits and various mimetic word labels annotated manually.

In this section, we introduce the procedure of the video recording session and the mimetic words labeling.

[1] http://www.murase.is.i.nagoya-u.ac.jp/~katoh/hoyo.html.

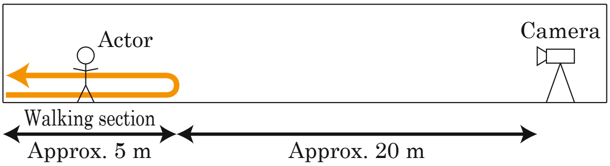


Fig. 1. Video recording environment.

Table 1. Selected mimetic words and their meanings [4,7].

Mimetic word	Meaning
suta-suta	Walk with light steps without observing around
noro-noro	Slowly walk without having a vigorous intention to move forward
yoro-yoro	Walk with an unstable balance
dossi-dossi	Walk with one's weight by stepping on the ground forcefully
seka-seka	Trot as being forced to hurry
teku-teku	Walk by firmly stepping on the ground for a long distance
tobo-tobo	Walk with dropping one's shoulder for a long distance
noshi-noshi	Walk with heavy steps forcefully
yota-yota	Walk with weak steps as with an elderly or a patient
bura-bura	Walk without having any intention

3.1 Video Recording

In this work, we use a kinetic feature following our previous work [4] as an input of the proposed framework. To collect kinetic coordinates, we detect body-parts (automatically detect and manually correct) from an image sequence captured from an ordinary camera, instead of using a depth sensor or a motion capture technique, because the mimetic words labeling procedure requires raw videos.

Figure 1 shows the environment of the video recording. The video recording was made over a single actor at a time. The walking section was approximately five meters long.

We asked ten amateur actors to walk with a gait representing a mimetic word back and forth the walking section. Here, the actors were native Japanese University students in their twenties but without professional acting skills. Table 1 shows a list of mimetic words instructed to the actors and their meanings, for reference. The ten mimetic words are commonly used ones, which were chosen from 56 mimetic words used to describe gaits listed in a Japanese mimetic word dictionary [7]. We asked the actors to walk with ordinary gaits as well. Finally, we recorded 292 gait videos (146 from the front of the actors and the paired 146 from their back).

The videos were taken at a rate of 60 fps, 527 × 708 pixels resolution, and 8-bit color. We used a USB 3.0 camera Flea3 produced by Point Gray Research,

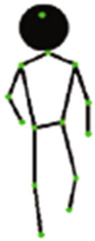


Fig. 2. Example of fourteen body parts.

Inc. The sensor size was 2/3 in., and the focal length of the lens was 35 mm. The camera was set approximately twenty meters away from the termination of the walking section. It aims to suppress the scale variation of body appearance due to walking along the optical axis of the camera.

3.2 Body-Parts Detection

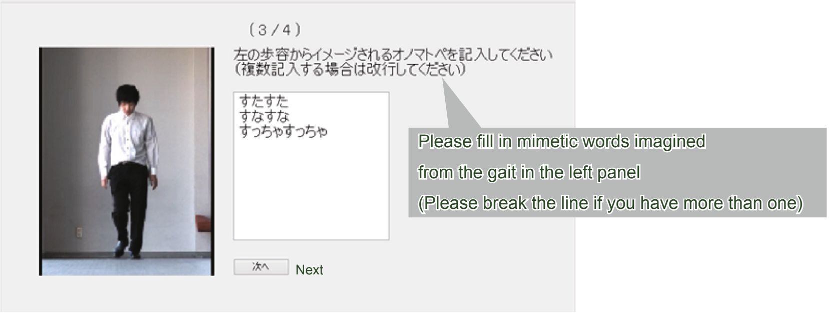
Li et al. proposed an algorithm for fine-grained classification of walking disorders arising from neuro-degenerative diseases such as Parkinson and Hemiplegia, by referring to relative body-parts movement [6]. In line with this work, we used kinetic features based on the relative movement of body parts in our previous research. To calculate them, we applied Convolutional Pose Machines (CPM) [14] to each frame of the dataset sequences mentioned above. Here, CPM is an articulated pose estimation method based on a deep learning model, which can detect fourteen parts of a human body, and yield their pixel coordinates.

However, the estimated body-parts coordinates are sometimes incorrect. In this paper, we use manually corrected data of the CPM detected coordination. For online applications, we will need a more accurate body-parts detector or a more convenient motion capturing device to obtain correct kinetic coordinates. Note that the dataset mentioned above includes the corrected body-part coordination data, and does not include the raw videos for the sake of the actors' privacies. Figures 2 shows an example of the fourteen body-parts.

3.3 Mimetic Words Labeling

In our previous work [4], the annotation was conducted in the form of choosing among ten types of candidates. Our framework has an ability of generating a variety of mimetic words, not only choosing one of the trained mimetic words. In order to make full use of the ability, the framework needs to learn various mimetic words, but the diversity of candidates was not enough in the previous work. To overcome this problem, in this work, we annotated more data, and also the annotators were allowed to give arbitrary mimetic words in a free description form.

Thirty annotators who are native Japanese University students in their twenties watched 146 videos showing the gaits from the front and annotated each



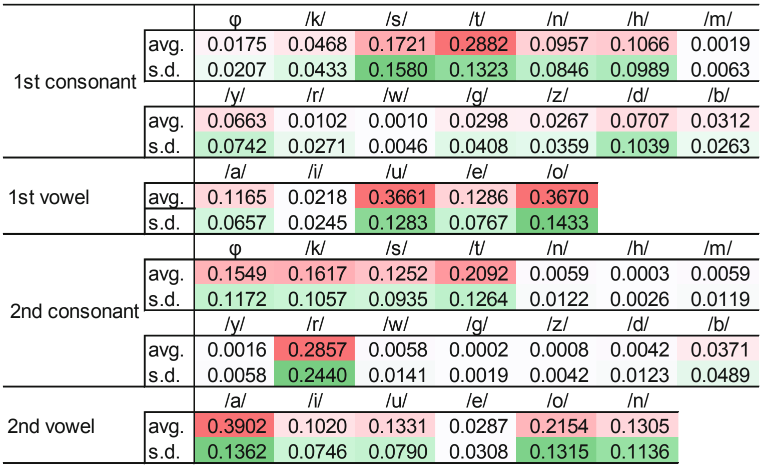
（３／４）

左の歩容からイメージされるオノマトペを記入してください
（複数記入する場合は改行してください）

すたすた
すなすな
すっちゃすっちゃ

Please fill in mimetic words imagined
from the gait in the left panel
(Please break the line if you have more than one)

Fig. 3. Annotation tool.



		φ	/k/	/s/	/t/	/n/	/h/	/m/
1st consonant	avg.	0.0175	0.0468	0.1721	0.2882	0.0957	0.1066	0.0019
	s.d.	0.0207	0.0433	0.1580	0.1323	0.0846	0.0989	0.0063
		/y/	/r/	/w/	/g/	/z/	/d/	/b/
	avg.	0.0663	0.0102	0.0010	0.0298	0.0267	0.0707	0.0312
	s.d.	0.0742	0.0271	0.0046	0.0408	0.0359	0.1039	0.0263
		/a/	/i/	/u/	/e/	/o/		
1st vowel	avg.	0.1165	0.0218	0.3661	0.1286	0.3670		
	s.d.	0.0657	0.0245	0.1283	0.0767	0.1433		
		φ	/k/	/s/	/t/	/n/	/h/	/m/
2nd consonant	avg.	0.1549	0.1617	0.1252	0.2092	0.0059	0.0003	0.0059
	s.d.	0.1172	0.1057	0.0935	0.1264	0.0122	0.0026	0.0119
		/y/	/r/	/w/	/g/	/z/	/d/	/b/
	avg.	0.0016	0.2857	0.0058	0.0002	0.0008	0.0042	0.0371
	s.d.	0.0058	0.2440	0.0141	0.0019	0.0042	0.0123	0.0489
		/a/	/i/	/u/	/e/	/o/	/n/	
2nd vowel	avg.	0.3902	0.1020	0.1331	0.0287	0.2154	0.1305	
	s.d.	0.1362	0.0746	0.0790	0.0308	0.1315	0.1136	

Fig. 4. Statistics of the freely described mimetic words.

video with three mimetic words they imagined. Fifteen annotators were assigned to each video, and annotated using the tool shown in Fig. 3. Here, the mimetic words were restricted to the pattern of *ABCD-ABCD*, which is the most common pattern of Japanese mimetic words. Note that *A* and *C* are consonants, and *B* and *D* are vowels.

Finally, 6,322 mimetic words were collected except for 248 invalid words, e.g. typing error or not in the *ABCD-ABCD* pattern. Statistics of the results are shown in Fig. 4. The upper row shows the mean occurrence frequency of each phoneme, and the lower row shows its standard deviation.

4 Gaits Description by Mimetic Words

The procedure of the proposed method based on our previous work [4] is shown in Fig. 5. In our method, we map the kinetic features extracted from videos

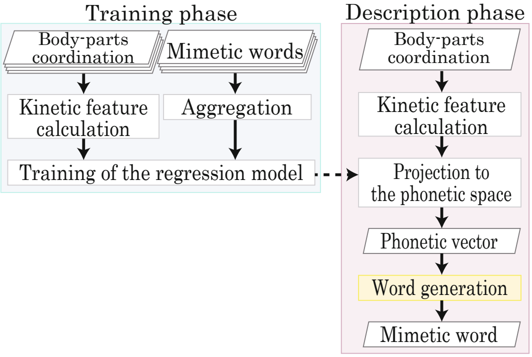


Fig. 5. Procedure of the proposed method.

to the phonetic space by regression. It consists of the training phase and the description phase. The main contribution of this paper is the proposal of the improved word generation module. Firstly, we explain the general framework concisely in Sect. 4.1. Secondly, the updated module is explained in Sect. 4.2.

4.1 General Framework of Describing Gaits

Takano et al. mentioned above showed the effectiveness of body-parts movement as a feature in describing gaits [13]. In addition, Li et al. proposed an algorithm for fine-grained classification of walking disorders arising from neurodegenerative diseases such as Parkinson and Hemiplegia, by referring to relative body-parts movement [6]. In line with these works, our framework [4] uses kinetic features based on the relative movement of body parts. Specifically, a sequence of arbitrary pairs of body-parts is used as an input.

Let the fourteen sequences of pixel coordinates be $\boldsymbol{P}(p,t) \in \mathbb{R}^2$. Here, $p \in \{0,\ldots,13\}$ indicates the index of each body part, and $t \in \{1,\ldots,T\}$ indicates the index of each video frame where the length of the input video is T [frames]. We calculate the Euclidean distance $D_{p_1,p_2}(t)$ between arbitrary pairs of parts p_1 and p_2. Then, we calculate the human height $H(t)$, namely, the difference in y-coordinates between head and foot, and their average in sequence $\bar{H}$. Finally, we divide all of $D_{p_1,p_2}(t)$ by $\bar{H}$, and obtain a sequence of the normalized body-parts distance $L_{p_1,p_2}(t)$. Note that the number of combinations of p_1 and p_2 under the condition of $p_1 < p_2$ is ${}_{14}\mathrm{C}_2 = 91$.

In order to handle mimetic words corresponding to gaits in a regression model, we express them in the form of "phonetic vector". As we mentioned in Sect. 3, in our dataset, multiple mimetic words can be annotated to each gait sequence. So we use the frequency vector of appearance of each phoneme composing the mimetic words corresponding to the gait as the phonetic vector $\mathbf{v}$. The vector is composed of 41 dimensions because the annotated mimetic words are restricted to the pattern of *ABCD-ABCD*, where A and C consist of fifteen consonants, B consists of five vowels, and D consists of six vowels [2]. Let the frequency vector of phonemes A, B, C, and D be $\mathbf{v}_A$, $\mathbf{v}_B$, $\mathbf{v}_C$, and $\mathbf{v}_D$ respectively, the phonetic vector $\mathbf{v}$ is represented as $(\mathbf{v}_A, \mathbf{v}_B, \mathbf{v}_C, \mathbf{v}_D)$. Note that $\mathbf{v}_A$, $\mathbf{v}_B$, $\mathbf{v}_C$, and $\mathbf{v}_D$ are normalized so that the summation of each element becomes 1.

Finally, a regression model learns the relation of the kinetic feature $L_{p_1,p_2}(t)$ and phonetic vector $\mathbf{v}$. Let the space constructed by the phonetic vector be named "phonetic space", the procedures can be regarded as estimating the mapping of the kinetic feature space to the phonetic space. In the description phase, the regression model estimates the phonetic vector $\hat{\mathbf{v}}$ from the kinetic feature $L_{p_1,p_2}(t)$.

4.2 Naturalness-Penalized Word Generation Module

This module generates an appropriate mimetic word from the estimated phonetic vector $\hat{\mathbf{v}}$ under consideration of "naturalness". Here, we define the co-occurrence frequency of phonemes composing a mimetic word as the naturalness.

Firstly, $\hat{\mathbf{v}}$ is split into the four frequency vectors for each phoneme; $\hat{\mathbf{v}_A}$, $\hat{\mathbf{v}_B}$, $\hat{\mathbf{v}_C}$, and $\hat{\mathbf{v}_D}$. This module chooses a mimetic word, i.e. series of phonemes, minimizing the following criteria.

$$\mathcal{L} = \mathcal{L}_d + \alpha \mathcal{L}_c \tag{1}$$

$$\mathcal{L}_d = ||\hat{\mathbf{v}_A} - Q(o_A)|| + ||\hat{\mathbf{v}_B} - Q(o_B)|| + ||\hat{\mathbf{v}_C} - Q(o_C)|| + ||\hat{\mathbf{v}_D} - Q(o_D)|| \tag{2}$$

$$\begin{aligned}\mathcal{L}_c = & w_{AB} C_{AB}(o_A, o_B) + w_{BC} C_{BC}(o_B, o_C) + w_{CD} C_{CD}(o_C, o_D) \\ & + w_{AC} C_{AC}(o_A, o_C) + w_{BD} C_{BD}(o_B, o_D) + w_{AD} C_{AD}(o_A, o_D)\end{aligned} \tag{3}$$

Here, each of o_A, o_B, o_C, and o_D is the candidate phoneme for each phoneme A, B, C, and D, respectively, and $Q(\cdot)$ is a function that converts a phoneme into a one-hot vector. $\mathcal{L}_d$ calculates the distance of $\hat{v}$ and mimetic words consisting of an arbitrary combination of phonemes. $\mathcal{L}_c$ is the naturalness penalty term. Note that $\mathcal{L}$ becomes an ordinary Nearest Neighbor method if the hyper parameter $\alpha = 0$, which corresponds to the previous method [4]. $C(\cdot)$ is the naturalness penalty between two phonemes.

For example, $C_{AB}(o_A, o_B)$ indicates a naturalness penalty of the first vowel o_A and the first consonant o_B. Finally, a combination of o_A, o_B, o_C, and o_D which minimizes the criterion $\mathcal{L}$ is obtained, and a mimetic word is output as the concatenation of the four phonemes.

The value of $C(\cdot)$ is calculated from the freely described mimetic words of the dataset introduced in Sect. 3.3. Firstly, all of annotated mimetic words are decomposed into a series of phonemes. Secondly, we aggregate these into histograms of each positional pair of phonemes, e.g. the first vowel and the first consonant. Thus, we obtain six $(= {}_4\mathrm{C}_2)$ co-occurrence histograms $C'(\cdot)$. Finally, the naturalness penalty $C(\cdot) = 1 - C'(\cdot)/N_{\mathrm{words}}$ is calculated per each positional pair. Here, N_{words} is the number of collected mimetic words (actually 6,322).

[2] In Japanese language, a special phoneme /n/ sometimes appears except in the first phoneme (it is called syllabic nasal). Although, strictly speaking, it is not a vowel, in this paper, we handle it as a vowel for convenience.

5 Experiments

We performed experiments for evaluating the correctness and the naturalness of the generated mimetic words. In Sect. 5.1, we report the result of a preliminary experiment to decide the weights of naturalness penalty w mentioned in Sect. 4.2. In Sects. 5.2 and 5.3, we report experiments evaluating the correctness and the naturalness of the generated mimetic words, respectively. Here, we define a subjective metric on how well a generated mimetic word expresses the corresponding gait as the correctness.

5.1 Parameter Tuning

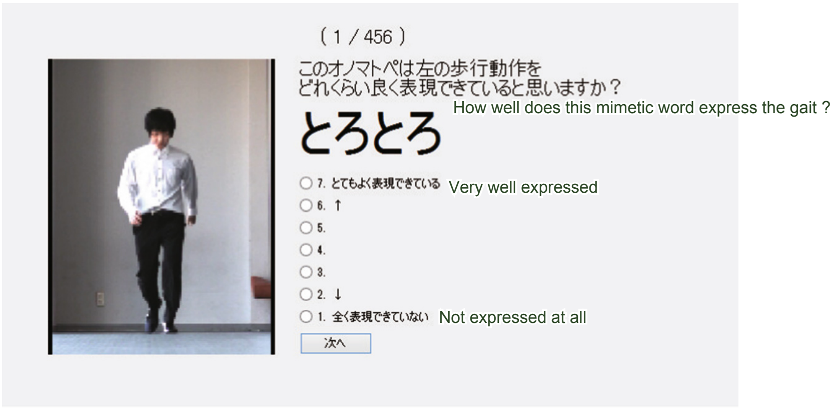
As we mentioned in Sect. 4.2, the proposed naturalness penalty criteria is composed of six terms. In this section, we report the result of a preliminary experiment to decide the weights of the penalty terms.

Firstly, we sorted all the 5,700 mimetic words generated from arbitrary combinations of phonemes according to the $\mathcal{L}_c$ criteria with equal weights. Secondly, we extracted ten words from the list of sorted mimetic words at an equal interval. Concretely, the following ten words were extracted: *yura-yura, guri-guri, zuze-zuze, sako-sako, done-done, maba-maba, roya-roya, pubu-pubu, hazo-hazo*, and *hape-hape*, in descending order. Thirdly, we conducted a pairwise comparison experiment for reranking the ten words into actual order of naturalness. We asked four evaluators to choose the more natural one from the pair of extracted words. The number of questions was ${}_{10}\mathrm{C}_2 = 45$. Then, we sorted the words in descending order of the selection rate. Finally, we grid-searched a combination of optimal weights. Each weight had a value of 0 to 9 with an increment of 1, and we calculated the naturalness ranking of the ten words under each condition. We searched the weights in which the calculated naturalness ranking had the highest correlation to the experimentally obtained actual naturalness ranking under Spearman's rank correlation criteria.

As a result, the following combination achieved the highest correlation 0.8389: $w_{AB} = 0$, $w_{BC} = 0$, $w_{CD} = 1$, $w_{AC} = 9$, $w_{BD} = 0$, $w_{AD} = 1$. Note that w_{AC} corresponds to the co-occurrence of the first consonant and the second consonant, w_{CD} corresponds to that of the second consonant and the second vowel, and w_{AD} corresponds to that of the first consonant and the second vowel. This result shows the importance of co-occurrence of two consonants. Incidentally, the most frequently appeared pair of consonants is the pair of the first consonant /t/ and the second consonant /k/. The words including this pair account for 797 words of all the collected 6,322 mimetic words through the annotation mentioned in Sect. 3.3. This pair often appears in popular mimetic words (e.g. "*toko-toko*" or "*teku-teku*"), and such a familiar combination of two consonants may take an important part in making us feel the mimetic word natural.

In the following experiments, we use this combination of weights. In other words, the naturalness penalty term becomes as follows:

$$\mathcal{L}_c = C_{CD}(o_C, o_D) + 9C_{AC}(o_A, o_C) + C_{AD}(o_A, o_D) \tag{4}$$



(1 / 456)

このオノマトペは左の歩行動作を
どれくらい良く表現できていると思いますか？

How well does this mimetic word express the gait ?

とろとろ

○ 7. とてもよく表現できている Very well expressed
○ 6. ↑
○ 5.
○ 4.
○ 3.
○ 2. ↓
○ 1. 全く表現できていない Not expressed at all

Fig. 6. User interface for correctness evaluation.

Table 2. Results of correctness evaluation.

Condition	Correctness (avg. ± s.d.)
$\alpha = 0$	4.434 ± 0.109
$\alpha = 1$	4.452 ± 0.088
$\alpha = 3$	4.275 ± 0.053
$\alpha = 6$	4.192 ± 0.067

5.2 Correctness Evaluation of the Description

In this section, we report an experiment for evaluating the correctness of the description.

We presented a pair of a gait video and a generated mimetic word to evaluators, and asked them how well the generated mimetic word described the gait from seven levels of Likert scale. Here, we call this metric as “correctness”. The presented gaits were the gait videos in the dataset introduced in Sect. 3, and the mimetic words were generated from phonetic vectors based on the freely described mimetic words for those videos. The evaluators were five native Japanese University students. Figure 6 shows the interface used for this evaluation. In this experiment, four methods were compared with hyperparameters $\alpha = 0, 1, 3$, and 6. As mentioned in Sect. 4.2, α is a parameter which decides the weight of the penalty term $\mathcal{L}_c$ to the distance $\mathcal{L}_d$, and when $\alpha = 0$, it becomes equivalent to the ordinary Nearest Neighbor method. The result is shown in Table 2. We can see that as α increases, the naturalness constraint becomes stronger. The correctness and naturalness are in the relation of a trade-off. The result shows that the condition $\alpha = 1$ can keep the correctness compared to the condition $\alpha = 0$. Note that the correctness evaluated under a random condition is 4.014. In the random condition, we presented a pair of a gait video and a random mimetic word to evaluators. Comparing these results, it was confirmed that the proposed method achieved higher correctness than the random description.

Table 3. Results of naturalness evaluation.

Condition	Naturalness (avg. ± s.d.)
$\alpha = 0$	4.962 ± 0.109
$\alpha = 1$	5.217 ± 0.077
$\alpha = 3$	5.356 ± 0.043
$\alpha = 6$	5.553 ± 0.052

5.3 Naturalness Evaluation of the Description

In this section, we report an experiment for evaluating the naturalness of the description.

We presented a generated mimetic word to evaluators, and asked them how natural the generated mimetic word is from seven levels of Likert scale. The evaluators were four native Japanese University students. As the same with the experiment in Sect. 5.2, four methods with $\alpha = 0, 1, 3$, and 6 were compared. The presented mimetic words were the same as in the previous experiment. The result is shown in Table 3. We can see that as α increases, the naturalness becomes higher.

Considering together with the evaluation result of correctness in Sect. 5.2, it turned out that the condition $\alpha = 1$ generates more natural mimetic words than the condition $\alpha = 0$ while maintaining the correctness.

6 Conclusions

In this paper, we proposed an improved mimetic word generation module considering naturalness, and updated our previously proposed description framework [4]. We defined the co-occurrence frequency of phonemes composing a mimetic word as the naturalness. We constructed a new dataset, and used the freely described mimetic words in the dataset to calculate the frequency. We formulated the naturalness penalty in six terms, each term corresponding to the co-occurrence of the positional pair of two phonemes. Through a preliminary experiment, we obtained the optimal weights of naturalness penalty terms, and revealed that the following three kinds of co-occurrences are important: the first consonant and the second consonant, the second consonant and the second vowel, the first consonant and the second vowel. To confirm the effectiveness of the proposed mimetic word generation module, we conducted two subjective experiments. Evaluators assessed the correctness and the naturalness in Likert scale. As a result, we confirmed that the proposed module could describe gaits with more natural mimetic words while maintaining the correctness.

Future works include exploring how the impression of human appearance (e.g. body shape or facial expression) biases a mimetic word we imagine.

Acknowledgements. Parts of this work were supported by MEXT, Grant-in-Aid for Scientific Research and the Kayamori Foundation of Information Science Advancement.

References

1. Doizaki, R., Watanabe, J., Sakamoto, M.: Automatic estimation of multidimensional ratings from a single sound-symbolic word and word-based visualization of tactile perceptual space. IEEE Trans. Haptics **10**(2), 173–182 (2017)
2. Fukusato, T., Morishima, S.: Automatic depiction of onomatopoeia in animation considering physical phenomena. In: Proceedings of the 7th ACM International Conference on Motion in Games, pp. 161–169 (2014)
3. Hamano, S.: The Sound-Symbolic System of Japanese. CSLI Publications, Stanford (1998)
4. Kato, H., et al.: Toward describing human gaits by onomatopoeias. In: Proceedings of the 2017 IEEE International Conference on Computer Vision, pp. 1573–1580 (2017)
5. Köhler, W.: Gestalt Psychology: An Introduction to New Concepts in Modern Psychology. WW Norton & Company, New York (1970)
6. Li, Q., et al.: Classification of gait anomalies from Kinect. Vis. Comput. **34**(2), 229–241 (2018)
7. Ono, M.: Jpn. Onomatopoeia Dict. (In Jpn.). Shogakukan Press, Tokyo (2007)
8. Ramachandran, V.S., Hubbard, E.M.: Synaesthesia–a window into perception, thought and language. J. Conscious. Stud. **8**(12), 3–34 (2001)
9. Sakata, A., Makihara, Y., Takemura, N., Muramatsu, D., Yagi, Y.: Gait-based age estimation using a DenseNet. In: Carneiro, G., You, S. (eds.) ACCV 2018. LNCS, vol. 11367, pp. 55–63. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-21074-8_5
10. Shimoda, W., Yanai, K.: A visual analysis on recognizability and discriminability of onomatopoeia words with DCNN features. In: Proceedings of the 2015 IEEE International Conference on Multimedia and Expo, pp. 1–6 (2015)
11. Sundaram, S., Narayanan, S.: Analysis of audio clustering using word descriptions. In: Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 769–772 (2007)
12. Sundaram, S., Narayanan, S.: Classification of sound clips by two schemes: using onomatopoeia and semantic labels. In: Proceedings of the 2008 IEEE International Conference on Multimedia and Expo, pp. 1341–1344 (2008)
13. Takano, W., Yamada, Y., Nakamura, Y.: Linking human motions and objects to language for synthesizing action sentences. Auton. Robot. **43**(4), 913–925 (2019)
14. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp. 4724–4732 (2016)