

Computational Measurement of Perceived Pointiness from Pronunciation

Chihaya Matsuhira^{1*}, Marc A. Kastner², Takahiro Komamizu³, Ichiro Ide¹, Takatsugu Hirayama⁴, Yasutomo Kawanishi⁵, Keisuke Doman⁶ and Daisuke Deguchi¹

^{1*}Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan.

²Graduate School of Informatics, Kyoto University, Yoshida Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan.

³Mathematical and Data Science Center, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan.

⁴Faculty of Environmental Science, University of Human Environments, 6-2 Kami Sanbonmatsu, Motojuku-cho, Okazaki, Aichi, 444-3505, Japan.

⁵Information R&D and Strategy Headquarters, RIKEN, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan.

⁶School of Engineering, Chukyo University, 101 Tokodachi, Kaizu-cho, Toyota, Aichi, 470-0393, Japan.

*Corresponding author(s). E-mail(s):

matsuhirac@cs.is.i.nagoya-u.ac.jp;

Contributing authors: mkastner@i.kyoto-u.ac.jp;

taka-coma@acm.org; ide@i.nagoya-u.ac.jp; t-hirayama@uhe.ac.jp;

yasutomo.kawanishi@riken.jp; kdoman@sist.chukyo-u.ac.jp;

ddeguchi@nagoya-u.jp;

Abstract

Sound symbolism is a well-researched topic of psycholinguistics, which tries to comprehend the connection between the sound of a word and its meanings. The *Bouba-Kiki effect*, one form of sound symbolism, claims that people perceive the pronunciation of “Kiki” as pointier than that

of “Bouba.” There is no research that focuses on modeling such perception, i.e., how pointy a pronunciation sounds to humans, through computational and data-driven approaches. To address this, this paper first proposes the novel concept of “phonetic pointiness” defined as how pointy a shape humans are most likely to associate with a given pronunciation. We then model this phonetic pointiness from computational and data-driven approaches to calculate a score for an arbitrary pronunciation. There are three proposed models: a referential model, an expressive model, and a combined model, which integrates the previous two. The idea comes from an existing psycholinguistic classification of two types of sound symbolisms: *referential symbolism* and *expressive symbolism*, where the former relates to vocabulary knowledge, while the latter is based on pure human intuition. The proposed models are constructed only with image and language data available on the Web, therefore not requiring task-specific human annotations. We evaluate these models through a crowd-sourced user study, finding a promising correlation between human perception and the phonetic pointiness calculated by the proposed models. The results indicate that human perception can be modeled better by combining both types of sound symbolisms. Furthermore, by observing the behaviors of the models, we show several possible use-cases, such as product naming and psycholinguistic research, which can be a useful insight to further studies and applications.

Keywords: Phonetic Impression, Pronunciation, Vision and Language, Image Generation, Affective Computing, Psycholinguistics

1 Introduction

Intuitive user interfaces are important for systems, tools, or applications to be useful and user-friendly. When designing complex user interfaces, understanding human perception is crucial. For this, a better understanding of how humans react to different stimuli is needed. For example, the knowledge that certain visual stimuli (such as the color red) or acoustic stimuli (such as beeping sounds) would alert people, can help design software user interfaces for warning people. Unlike the perception towards such stimuli, the perception towards languages, especially those related to the pronunciation of words, has received scarce attention. Therefore, in this research, we tackle the research question of what we perceive from a word pronunciation, such as “what kind of word pronunciation would alert people the most out of synonyms.”

Sound symbolism [1] is a well-researched topic of psycholinguistics, which states that the pronunciation of a word and its meaning are not fully arbitrary. One form of sound symbolism can be observed in the *Bouba-Kiki effect* [2, 3]. Figure 1 shows two shapes of contrasting pointy and round features. According to the phenomenon, if asked which shape is called *Bouba* and which is called *Kiki*, people tend to associate the round shape (right) with *Bouba* and the pointy shape (left) with *Kiki*.

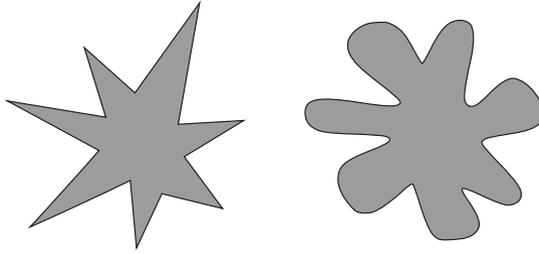


Fig. 1: Two shapes often used to describe Bouba-Kiki effect¹. If asked which shape is called *Bouba* and which is called *Kiki*, people tend to associate the round shape (right) with *Bouba* and the pointy shape (left) with *Kiki*.

As a first step to modeling the general sound symbolism, this paper focuses on the perceived pointiness from pronunciation, which the Bouba-Kiki effect represents. We first propose and introduce the concept of “phonetic pointiness,” a type of impression humans perceive from pronunciation. We define it as how pointy a shape humans are most likely to associate with a given pronunciation. Next, we model the phonetic pointiness through computational and data-driven approaches. According to the definition, the phonetic pointiness should deeply relate to the Bouba-Kiki effect, and thus the knowledge of this phenomenon can be useful for modeling it. Hence, the rest of this paper will focus on measuring the human reaction towards this phonetic pointiness by making use of the psycholinguistic findings about sound symbolism including the Bouba-Kiki effect, as well as techniques from Computer Vision (CV) and Natural Language Processing (NLP).

In the English language, only a few studies have been made to compute phonetic impressions as well as human perception towards them. To the best of our knowledge, no study exists that focuses on the concept of phonetic pointiness.

As related work, K. Papantoniou and S. Konstantopoulos [4] predict the polarity of a character’s role from his or her name by training a model with manually annotated data. V. Sabbatino et al. [5] estimate which emotion a pseudo-word would evoke in humans by using emotion labels annotated to existing words. Their assumption for expanding the emotion labels of existing words to pseudo-words is that impressions perceived from a pseudo-word (e.g., “snice”) would be influenced by the impressions of existing words having similar pronunciation (e.g., “nice”).

This assumption has also been studied in psycholinguistics. E. Sapir [6] proposes to classify sound symbolism into two types: *referential symbolism* and *expressive symbolism*. The former is caused by human knowledge of words, while the latter is caused by human intuition. Both types are explained in Fig. 2 with an example English pseudo-word /'kivi/ (pronounced as Kee-Vee). When we hear the pronunciation /'kivi/, we may imagine a similar-sounding

¹Authors: Bendž, Qef, and Andrew Dunn, CC Attribution-Share Alike 3.0, from <https://commons.wikimedia.org/w/index.php?curid=19653163>.

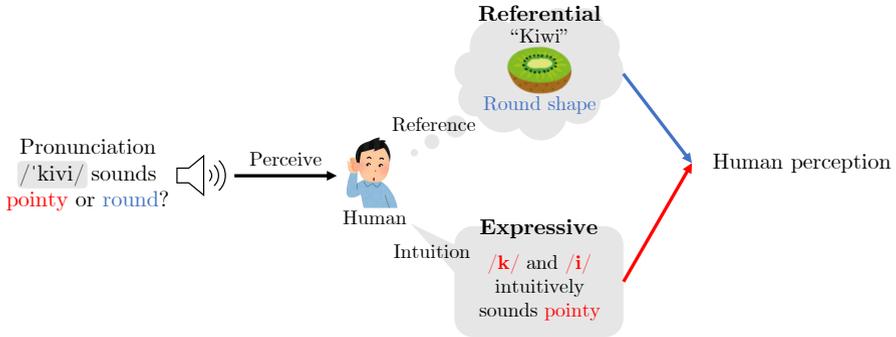


Fig. 2: Concept of referential symbolism and expressive symbolism by E. Sapir [6] in answering phonetic pointiness of pronunciation.

word “Kiwi” (/’kiwi/), which could evoke a round impression from the round shape of “Kiwifruits.” Meanwhile, we may perceive the sound /’kivi/ pointy, intuitively from its pronunciation, like “Kiki” in the Bouba-Kiki effect. In this example, the former impression can be regarded as the referential symbolism, and the latter can be regarded as the expressive symbolism. In addition to this theory, this paper assumes that both types of symbolisms are mixed to form the full human perception, as illustrated in the right part of Fig. 2.

Based on these ideas, this paper proposes two different types of models, a referential symbolism model (in short, *referential model*) and an expressive symbolism model (in short, *expressive model*), to calculate the phonetic pointiness for an arbitrary pronunciation written with International Phonetic Alphabet (IPA) [7]. Both models take the pronunciation of a word as an input and output its calculated phonetic pointiness score, yet their targets are different. The referential model first refers to an existing word (e.g., “Kiwi” (/’kiwi/)) phonetically similar to an input pronunciation (e.g., /’kivi/). Next, the visual pointiness of its shape is measured to be the final output of the model. In contrast, the expressive model attempts to model pure human intuition. To achieve this, we take advantage of existing findings of the Bouba-Kiki effect in psycholinguistics. Specifically, we construct the model based on the idea that the output of the model should be low when an input pronunciation is similar to Bouba, and high when it is similar to Kiki. Lastly, we attempt to construct a combined model, which combines the referential and expressive models. As explained in the previous paragraph, since human reaction towards phonetic pointiness is assumed to be influenced by both referential and expressive symbolisms, the combined model is expected to fit human perception than the individual models better.

Constructing these three proposed models does not require task-specific human annotations unlike the previous studies [4, 5]. Instead, we leverage image and language data available on the Web, thus making the training less expensive. Moreover, this can provide a benchmark of how well human perception can be modeled without human annotations.

The performance of the proposed models is evaluated through user study via crowd-sourcing. We create 400 random pseudo-words, which are used in asking participants to choose which of the two pseudo-words sounds rounder or pointier. In this paper, we measure how accurate the proposed models can estimate their answers, and discuss their performance based on the accuracy.

Because this research proposes methods to calculate the phonetic pointiness scores for arbitrary pronunciations, various applications can be considered. For example, the proposed models can be used to assist people in naming commercial products. Other applications include improving human-computer interaction interfaces and helping machines understand human perception towards languages. Furthermore, by constructing models in different languages and comparing them, this research could be a tool to be used in psycholinguistics, comparative linguistics, cultural sociology, and even etymology.

In these applications, not only the optimal combined model but also the referential model and the expressive model can be useful on their own, depending on the contexts. The referential model can assess what kinds of existing words humans would associate with, while the expressive model can predict how humans would intuitively perceive when hearing a pronunciation. As such, studies using our models have a choice of which model to use depending on their use-cases.

In summary, our main contributions are:

1. the introduction of the novel concept of phonetic pointiness,
2. the first attempt to model the phonetic pointiness by means of computational and data-driven approaches, and
3. modeling, comparing, and analyzing two different aspects of sound symbolism, i.e., the referential aspect and the expressive aspect, to find out that the combination of both symbolisms explains human perception better.

This paper is structured as follows. In Section 2, psycholinguistic research of the Bouba-Kiki effect is first introduced, followed by an introduction of computational approaches to modeling sound symbolism. Section 3 proposes three models to calculate phonetic pointiness, and experiments conducted to evaluate the proposed models are reported in Section 4. Lastly, Section 5 discusses applicability of the models before concluding the paper in Section 6.

2 Related Work

To date, no work has achieved our goal of computing phonetic pointiness scores for arbitrary pronunciations. Yet, to proceed with our research, it is essential to recognize the psycholinguistic findings of the Bouba-Kiki effect and how previous computational approaches have modeled sound symbolism. In this section, Section 2.1 first introduces several psycholinguistic studies of the Bouba-Kiki effect. Next, Section 2.2 introduces the current computational approaches to sound symbolism.

2.1 Bouba-Kiki Effect

The Bouba-Kiki effect was first discovered in 1929 by a psychologist W. Köhler [2]. He showed two round/pointy shapes similar to Fig. 1 to inhabitants on Tenerife island, Spain, and asked them which shape has a name *Maluma* and which shape has a name *Takete*. As a result, he confirmed that most of the inhabitants associated the round one with *Maluma* and the pointy one with *Takete*. Afterwards, in 2001, V.S. Ramachandran and E.M. Hubbard [3] repeated Köhler's experiment using alternative pseudo-words *Bouba* and *Kiki* instead of *Maluma* and *Takete*, and reproduced the result as participants tended to associate the round shape with *Bouba*, and the pointy shape with *Kiki*.

Since then, there has been a lot of psycholinguistic research on this phenomenon. Several researchers conducted Bouba-Kiki experiments on children as well as adults and compared their reactions [8, 9]. They confirmed that children, and even infants who have not fully acquired their languages yet, showed the same Bouba-Kiki preferences as adults, indicating the possibility that such human perception towards the Bouba-Kiki effect precedes children's language acquisition.

Others explored how much the cultural backgrounds of participants influence the Bouba-Kiki effect. A.J. Bremner et al. [10] conducted experiments on Western and African people, and compared the difference in their preferences for both shape-sound and shape-flavor matching. The former experiment was the same as Köhler's, asking participants to match two different pronunciations with round/pointy shapes. In the latter experiment, on the other hand, they were asked to match two flavors of still and sparkling water with round/pointy shapes. As a result, in the shape-sound matching, both groups showed the same tendency of associating Bouba with the round and Kiki with the pointy shape. In the shape-flavor matching, however, Western people tended to associate sparkling water with the pointy shape, while African people tended to associate the opposite way, suggesting that there might be a cultural difference in Bouba-Kiki preferences, especially in the shape-flavor matching. Moreover, Y.C. Chen et al. [11] focused on the difference in preferences in shape-sound matching among North American and Taiwanese people. They found out that the overall tendency of their preferences towards the Bouba-Kiki effect is consistent, yet it is partly tuned by their cultural backgrounds. For example, what visual characteristics are considered important may differ among groups from different cultural backgrounds. More recently, Ówiek et al. [12] have shown evidence for the Bouba-Kiki effect being robust across languages and writing systems by recruiting speakers of 25 different major languages. According to their work, human preferences towards the Bouba-Kiki effect are not strongly affected by the language knowledge of the participants and the shapes of the alphabets of their native languages.

These studies suggest that, in general, the Bouba-Kiki effect is independent of the ages, cultural backgrounds, and native languages of participants. Despite this fact, S.J. Styles and L. Gawne [13] analyzed under which conditions the

Bouba-Kiki effect can or cannot be observed. They focused on a case reported by S.K. Rogers and A.S. Ross [14] in which the Bouba-Kiki effect cannot be observed and analyzed the reason for this. According to their results, eventually, it was shown that the characteristic of how natural pseudo-words sound to participants plays an important role in this problem. In other words, if the occurrence of the pronunciation of Bouba (i.e., written with IPA as /'bubɑ/) is not allowed by the phonology of a participant's native language, the participant may not show the Bouba-Kiki correspondence in that experiment. Therefore, when conducting experiments similar to these, the pseudo-words to be shown to participants should be carefully selected based on the phonology of the participants' native languages.

Some researchers conducted experiments with new pseudo-words other than Bouba, Maluma, Kiki, and Takete, and attempted to observe how humans react towards different kinds of pronunciations [15, 16]. In these studies, the format of pseudo-words used in their experiments is restricted to CVCV (Consonant+Vowel+Consonant+Vowel). Since the original pseudo-words Bouba (e.g., /'bubɑ/) and Kiki (e.g., /'kiki/) by V.S. Ramachandran and E.M. Hubbard had such a format, it is a reasonable first step to focus only on the CVCV format pseudo-words.

Others attempted to unveil the pointiness associations of phonemes [15, 17, 18]. According to their findings, open back vowels (/ɑ/, /ɔ/) and rounded vowels (/u/, /o/) tend to be associated with round shapes, and close front unrounded vowels (/i/, /e/) tend to be associated with pointy shapes. As for consonants, sonorants (/m/, /l/) and voiced consonants /b/, /g/ tend to be associated with round shapes, and voiceless plosives (/p/, /k/, /t/) tend to be associated with pointy shapes.

Although these psycholinguistic findings point out several cardinal characteristics of pronunciation that can lead to the estimation of the phonetic pointiness, they are still not sufficient to measure the concept for an arbitrary pronunciation. Moreover, all the studies introduced in this section mostly focused on the expressive aspect of sound symbolism, hence not considering the referential aspect.

2.2 Computational Models for Sound Symbolism

Regarding the English language, only a few studies [4, 5] exist that aim to model sound symbolism including the Bouba-Kiki effect using computational and data-driven approaches. Studies in this direction include predicting the polarity of a character's role in a movie from his or her name [4], and estimating which emotion a pseudo-word would evoke [5]. The former study first obtains human annotations via crowd-sourcing, and constructs a regression model to predict the polarity with the manually annotated data. The input of the model is mainly a set of handmade phonological features, such as the number of certain phonemes in the input. Besides, since the authors are also interested in what kinds of name attributes relate to the polarity, they also input other types of features, such as the gender of the character and the genre of the

movie. The latter study focuses on estimating emotion labels evoked by a pseudo-word by using emotion labels annotated on existing words. To obtain a mapping between them, the authors train a language model with character- or phoneme-level embeddings. In their study, they expand attributes (emotion labels) of existing words to pseudo-words by considering the phonetic similarity between existing words and the target pseudo-word. For example, they assume that a pseudo-word “snice” would evoke a positive impression because of the similar-sounding existing word “nice.” This idea is the same as the one used to construct the referential model in our study.

In both studies, unlike ours, the phonetic pointiness is not the main focus of the research. Moreover, they train their model using manually annotated data, which requires a huge number of human annotations. Furthermore, we notice a problem that the names and pseudo-words used in the studies are mostly written with alphabets (e.g., “phrouth”). In this case, certain tools are required to automatically convert the spellings to phonetic symbols. However, this may cause ambiguity, since the pronunciation deduced from the spelling is not always unique in the English language. For example, the spelling “ou” may be pronounced as /aʊ/ (as in “mouth”), but may also be /u/ (as in “youth”), /oʊ/ (as in “soul”), or /ɔ/ (as in “bought”). In our research, in contrast, pseudo-words are generated on the basis of phonetic symbols (e.g., /'fɹaʊθ/ instead of “phrouth”), thus avoiding this ambiguity.

While modeling sound symbolism is still a developing field, sound symbolism itself has been actively used to improve user interfaces of systems. For most of these existing studies, the main research interest has focused on onomatopoeia. Onomatopoeia is a type of sound symbolism by which a state, a sound, or a movement of a target is depicted by specific pronunciation. For example, the English words “Tick-Tock” (sound of a clock) and “Bang” (sound of explosion) can be regarded as onomatopoeia words. Other examples include animal cries such as “Bow-Wow” for dogs and “Meow” for cats. These onomatopoeia words are especially common in comics and advertisements since they can express sounds and movements only with text descriptions.

O. Gillet and G. Richard [19] integrated onomatopoeia information into a drum loop retrieval system. They used onomatopoeias of the sounds of drum loops such as [ts] (for hi-hat) and [tʃa] (for snare drum) to improve the retrieval performance of their system. S. Sundaram and S. Narayanan [20] combined onomatopoeia with audio classification. They improved the applicability of a classification model by adding onomatopoeia information to its training labels. In detail, they used onomatopoeia labels like “Tap” and “Clatter” as well as general classification labels like “Doors” and “Office,” and trained a model that can output such classification labels for input audio data. T. Fukusato and S. Morishima [21] proposed a method to automatically select an appropriate onomatopoeia word for a certain CG animation. By referring to physical parameters of target objects such as mass and acceleration, they estimated the most appropriate onomatopoeia words including “Bong” and “Zap” for a collision scene between two objects.

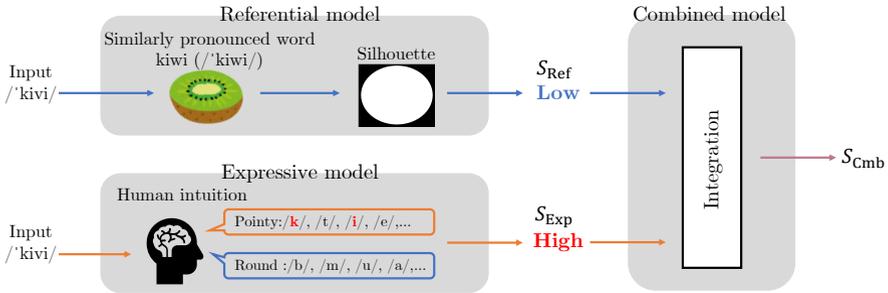


Fig. 3: Overview of the proposed models to calculate phonetic pointiness from pronunciation.

Some onomatopoeia words are existing words, while some can be regarded as pseudo-words. Due to this, attempts to analyze the human perception towards pseudo-words like ours, could foster the understanding of onomatopoeia and also these studies.

3 Calculation of Phonetic Pointiness

In this paper, we propose three models to calculate phonetic pointiness from an input pronunciation: a referential symbolism model, an expressive symbolism model, and a combined model of the previous two.

The overview of these models is shown in Fig. 3. The referential model first refers to images of existing words that are pronounced similarly to an input pronunciation (e.g., images of a *kiwi*(fruit) for an input /kivi/). Next, after extracting the silhouette of the object, it measures the visual pointiness of the silhouette (how pointy the silhouette looks) to obtain the final score S_{Ref} . If the silhouette shape of the object is not pointy, S_{Ref} would be low, and if it is pointy, S_{Ref} would be high. Meanwhile, the expressive model outputs a score S_{Exp} in terms of how pointy an input pronunciation intuitively sounds to humans. If the input pronunciation is similar to *Bouba* or *Maluma*, S_{Exp} would be low, and if it is similar to *Kiki* or *Takete*, S_{Exp} would be high. The combined model uses the output scores of these two models, and calculates a combined score S_{Cmb} .

3.1 Referential Symbolism Model

The procedure of calculating the phonetic pointiness using the proposed referential symbolism model (in short, *referential model*) is shown in Fig. 4a. The referential model consists of two steps: silhouette image generation from an input pronunciation and visual pointiness calculation from the generated image.

From an input, the image generator first generates an image that displays an object silhouette of a word, or a mixture of silhouettes of words, which are pronounced similarly to the input. For example, if the input is /kivi/,

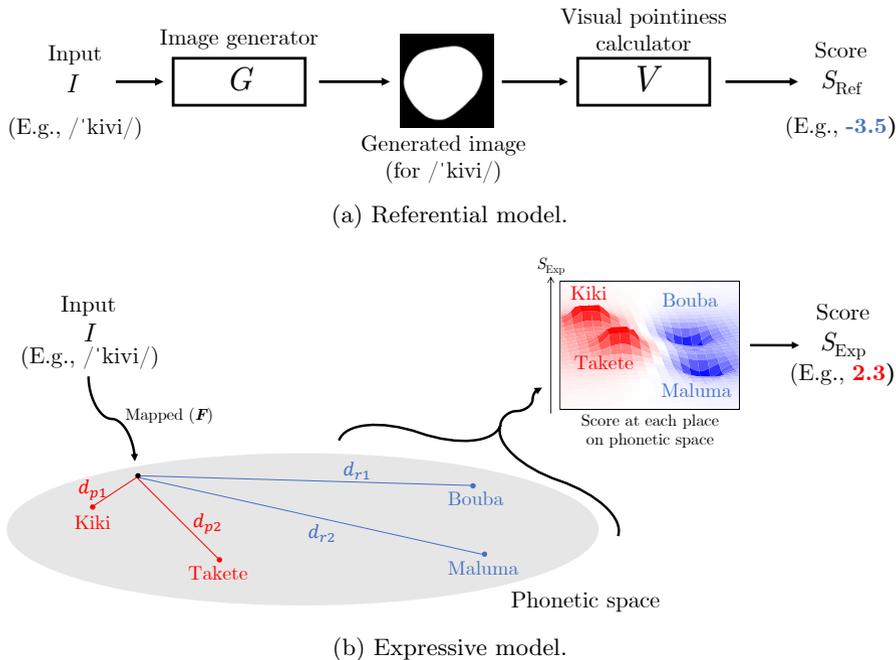


Fig. 4: Detailed procedure of calculating phonetic pointiness by the proposed models.

the generator is expected to output a silhouette image similar to a “Kiwi” (Kiwifruit), which is phonetically similar to /'kivi/. Next, the visual pointiness of the silhouette is measured from the generated silhouette image. By doing this, the output score can reflect the visual pointiness of the shapes of similar-sounding words to the input pronunciation.

As Fig. 4a illustrates, the formulation of the referential model can be written as

$$S_{\text{Ref}} = V(G(I)), \quad (1)$$

where I is an input pronunciation, and G and V represent the image generator and the visual pointiness calculator, respectively.

3.1.1 Image Generation from Pronunciation

For the image generator of the referential model, we adopt *Deep Fusion Generative Adversarial Network* (DF-GAN) [22]. DF-GAN is a type of Generative Adversarial Network (GAN) [23] and it works as a text-to-image synthesis model; It takes a caption as an input and generates an image that matches the input caption. It has three stacked image generators. The first generator generates 64×64 pixels from a latent vector concatenated with a conditional vector. The second and the third ones generate a higher resolution image from the image generated by the previous generator. As a result, the resolution of

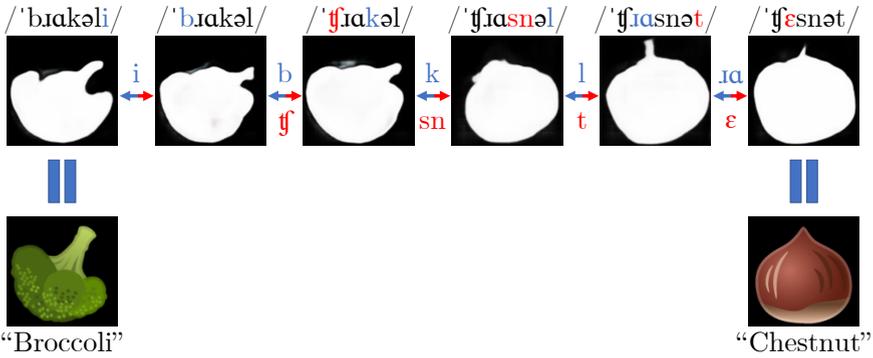


Fig. 5: Examples of silhouette images generated by the image generator trained on our vegetable dataset, illustrating the concept that the more similar the input pronunciation gets to an existing word, the more the shape of the silhouette resembles the shape of the word.

the output of the DF-GAN, i.e., the image generated by the third generator, becomes 256×256 pixels.

More technically, the input of the original DF-GAN is an array of words, such as [“a”, “blue”, “bird”, “is”, “flying”]. In the referential model, on the other hand, the input must be an arbitrary pronunciation. To bridge this gap, we regard each pronunciation as an array of phonemes, thus the DF-GAN can generate an image from an array of phonemes. For example, if a pronunciation /kivi/ is input, we feed the DF-GAN with an array of phonemes [/, /k/, /i/, /w/, /i/]. Next, in the same way as the original DF-GAN calculates, each phoneme is embedded by a Pronunciation-Encoder (called Text-Encoder in reference [22]), and then the embeddings are used as a conditional vector from which an output image is generated. In the training phase, for an existing word “Kiwi” (pronounced as /kiwi/), for example, we feed the DF-GAN with an array of phonemes [/, /k/, /i/, /w/, /i/] as an input, and a silhouette image of a kiwi as an expected output. By training the image generator in this way, we can obtain silhouette images for an arbitrary pronunciation that reflect the silhouette shapes of similar-sounding words.

Examples of the silhouette images generated by the image generator trained on vegetable names and images (the dataset will be described in Section 4) are shown in Fig. 5. As expected, the more similar the input pronunciation gets to an existing word (e.g., “Broccoli” and “Chestnut”), the more the shape of the silhouette resembles the shape of the word.

3.1.2 Visual Pointiness Calculation

For the visual pointiness calculator, we adopt a metric established in our previous study [24]. The metric measures how pointy an image is perceived by humans based on the pointiness/roundness of objects appearing in the image.

For example, the metric shows a high score for a silhouette image of a spiky-shaped object, and a low score for a silhouette image of a circle. The metric has robustness against several visual features: scale, position, rotation, occlusion, and number of objects in an image. In addition, the metric shows a lower score when the silhouette of an object is blurred.

The calculation of this metric comprises four steps: 1. feature map calculation, 2. histogram calculation, 3. histogram standardization, and 4. summation calculation. First, we calculate a feature map from an image, which describes the degree of visual pointedness for each pixel. To obtain this, for each pixel, we count the number of continuous surrounding pixels that have a darker intensity than that of the target pixel. The number at each pixel becomes an integer ranging between 0 and 16. Next, we calculate a histogram from the feature map. Each pixel of the map is an integer from 0 to 16. Our previous study assumed that only the numbers between 9 and 15 correlate with the visual pointiness at the target pixel well. Hence, we make a histogram with seven bins. A value of the n -th bin of the histogram y_n is calculated as,

$$y_n = \frac{x_n}{\sum_{n \in \{9, \dots, 15\}} x_n}, \quad (2)$$

where x_n is a number of pixels having a value n in the feature map. Then, we standardize the histogram. With an estimated mean $\hat{\mu}_n$ and an estimated standard deviation $\hat{\sigma}_n$, we calculate the value $\frac{y_n - \hat{\mu}_n}{\hat{\sigma}_n}$ for each bin. For obtaining the estimated values, we use unbiased estimators calculated as,

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^m y_{ni}, \quad \hat{\sigma}_n = \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} \sqrt{\frac{1}{2} \sum_{i=1}^m (y_{ni} - \hat{\mu}_n)^2}, \quad (3)$$

where y_{ni} is the y_n for the i -th image and $\Gamma(x)$ is the gamma function. Finally, the visual pointiness score, i.e., the output of the referential model S_{Ref} , is calculated by summing all values of the standardized histogram, which is

$$S_{\text{Ref}} = \sum_{n \in \{9, \dots, 15\}} \frac{y_n - \hat{\mu}_n}{\hat{\sigma}_n}. \quad (4)$$

Through the calculation described above, we add a slight modification to the original implementation. It is in the fourth step, where the method above calculates a summation of all values of the histogram. Instead of a summation, the original one calculates a weighted sum with a hand-made weight function. The intention of the weighted sum is to make sure that pointy edges (e.g., pixels where $n = 15$) are treated as pointier than less pointy ones (e.g., pixels where $n = 9$). However, as such information is already retained through step 3, i.e., the preceding histogram standardization in the metric calculation, we replace the weighted sum calculation with just a summation in this paper.

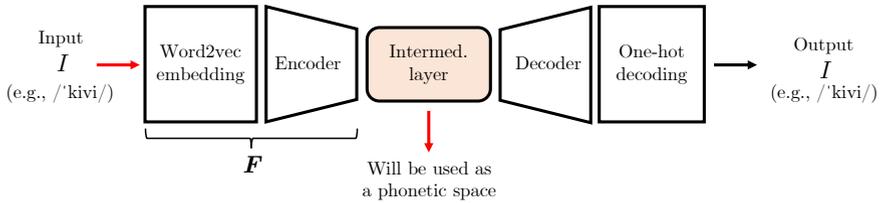


Fig. 6: Architecture of Auto-Encoder used to construct a phonetic space.

3.2 Expressive Symbolism Model

The procedure of calculating the phonetic pointiness by the expressive symbolism model (in short, *expressive model*) is shown in Fig. 4b. The expressive model consists of two steps: phonetic space construction and score calculation on the space. The core assumption is that the output score should be low if the input pronunciation is similar to Bouba or Maluma, and high if the input pronunciation is similar to Kiki or Takete, since (*Bouba*, *Kiki*) and (*Maluma*, *Takete*) are the most notable examples of contrasting pronunciation pairs used to confirm the Bouba-Kiki effect.

To calculate the score, we define two pronunciation sets W_r and W_p in advance, each of which is composed of certain words that are perceived as round or pointy by humans. According to the findings of the Bouba-Kiki effect described in Section 2, for example, W_r should contain Bouba and Maluma, while W_p should contain Kiki and Takete. The actual combinations of W_r and W_p tested in our experiments will be described in Section 4.

In the score calculation, the input pronunciation is first mapped onto a phonetic space as well as every word in W_r and W_p , where similar-sounding words will be located close to each other. Next, the similarity between the input pronunciation and each word in W_r and W_p are measured on the phonetic space, and the final output is calculated based on the similarities. Note that, throughout the calculation of this expressive model, we attempt to model human intuition without using human annotations.

3.2.1 Phonetic Space Construction

The phonetic space of the expressive model is constructed using an Auto-Encoder architecture, as shown in Fig. 6. Each phoneme in an input pronunciation is first encoded into a phoneme embedding by Word2vec [25, 26]. Word2vec originally is a model for encoding words into word embeddings. In the same scheme as we did with DF-GAN explained in Section 3.1.1, we extend the idea of Word2vec word embeddings to phoneme embeddings. In detail, an input pronunciation (e.g., /'kivi/) is first regarded as an array of phonemes (e.g., [i/, /k/, /i/, /v/, /i/]), and then Word2vec is trained to estimate the adjacent phonemes for each phoneme in the array. After pre-training Word2vec with a number of pronunciations of existing words, the pre-trained

one is used to obtain phoneme embeddings. Using such Word2vec embeddings is more effective than using simpler encodings such as One-hot encoding, since Word2vec can learn the implicit phonological relationship among phonemes from occurrences of phonemes in pronunciations of existing words [27].

The phoneme embeddings obtained by the Word2vec are then input to an encoder that employs Long Short-Term Memory (LSTM) [28], and the encoded vector is mapped onto its intermediate representation, which is the phonetic space. This LSTM-based encoder can consider sequential information of input pronunciations, and thus the encoder can distinguish an English word “Tap” (/ˈtæp/) from another English word “Pat” (/ˈpæt/). Finally, an LSTM-based decoder decodes the vector on the phonetic space to retrieve its corresponding original pronunciation. Unlike the encoding process, this decoding process does not compute Word2vec decoding.

The encoder and decoder are simultaneously trained as an Auto-Encoder using pronunciation data of existing words. Its intermediate representation is used as a phonetic space in the following steps.

3.2.2 Score Calculation on Phonetic Space

Given an input pronunciation I and two pronunciation sets $W_r = \{w_{r_1}, w_{r_2}, \dots\}$ and $W_p = \{w_{p_1}, w_{p_2}, \dots\}$, dissimilarities (Euclidean distances in this paper) among their pronunciations are first measured on the phonetic space obtained through the process described in Section 3.2.1. Next, based on these dissimilarities, the final score S_{Exp} is calculated as,

$$S_{\text{Exp}} = \sum_{w_p \in W_p} \frac{1}{d(\mathbf{F}(I), \mathbf{F}(w_p))} - \sum_{w_r \in W_r} \frac{1}{d(\mathbf{F}(I), \mathbf{F}(w_r))}, \quad (5)$$

where \mathbf{F} represents the encoder of our Auto-Encoder depicted in Fig. 6, which maps pronunciations to vectors on the phonetic space, and $d(\mathbf{x}, \mathbf{y})$ represents the Euclidean distance between two vectors \mathbf{x} and \mathbf{y} .

According to this formula, the closer the input I approaches to a round word w_r , the lower the score S_{Exp} will be, and the closer I approaches to a pointy word w_p , the higher S_{Exp} will be. Note that S_{Exp} will be $-\infty$ if I is identical to the round word w_r , and S_{Exp} will be $+\infty$ if I is identical to the pointy word w_p .

3.3 Combined Model

The combined model takes two scores S_{Ref} and S_{Exp} as an input, and calculate the combined score S_{Cmb} . First, variances of S_{Ref} and S_{Exp} are normalized to be \hat{S}_{Ref} and \hat{S}_{Exp} which are calculated as,

$$\hat{S}_{\text{Ref}} = \frac{S_{\text{Ref}}}{\sigma_{\text{Ref}}}, \quad \hat{S}_{\text{Exp}} = \frac{S_{\text{Exp}}}{\sigma_{\text{Exp}}}, \quad (6)$$

where σ_{Ref} and σ_{Exp} represent the standard deviations of S_{Ref} and S_{Exp} , respectively. Next, using a weighting parameter a , the score S_{Cmb} is calculated as,

$$S_{\text{Cmb}} = a\hat{S}_{\text{Ref}} + (1 - a)\hat{S}_{\text{Exp}}. \quad (7)$$

Since \hat{S}_{Ref} and \hat{S}_{Exp} are normalized to have equal variance, the weight a can be interpreted as the ratio of the importance of the outputs of both models. If $a = 0$, the output of the combined model will be substantially equivalent to that of the expressive model. If $a = 1$, it will be equivalent to that of the referential model. For our further experimentation and analysis, we need to decide the optimal weight a that best matches human perception. Because such a weight is not obvious, in the experiments conducted in Section 4, we apply grid-search for searching the optimal a for each experiment. We further attempt to reveal which symbolism is more important in which condition and to what extent the performance will improve if we combine two types of models.

4 Experiment

We conduct two experiments to evaluate the proposed models on a task of estimating the human preference of the phonetic pointiness towards a pair of pseudo-words. The purposes of these experiments are: (1) to evaluate and compare how well each proposed model grasps human perception, and (2) to identify in what kind of pronunciation each model/symbolism has strengths and weaknesses. For these purposes, we collect participants' responses through user study via crowd-sourcing. Although the Bouba-Kiki effect and the proposed phonetic pointiness are believed to be independent of participants' mother languages, this paper restricts the target only to the English language.

In these experiments, we focus on two different pronunciation features: phoneme types in pronunciation and pronunciation length. Accordingly, the first experiment is set to measure the performance towards types of phonemes (vowels or consonants) in a simple setting, while the second experiment measures the performance towards pronunciation length in a more complicated setting.

These two experiments only differ in the pseudo-words used in each of the experiments. The experiment on phoneme types (in short, *phoneme experiment*) uses pseudo-words having a pronunciation in CVCV (Consonant+Vowel+Consonant+Vowel) format. This follows the precedent studies of psycholinguistics [15, 16]. Additionally, pronunciations of two pseudo-words in each question are set similar (e.g., /'bubav/ and /'wuwav/). These restrictions are set with the aim of making questions easier for participants to respond to. The experiment on pronunciation length (in short, *length experiment*) uses more arbitrary pseudo-words in terms of pronunciation length. Each question consists of two pseudo-words having a pronunciation in either CVC, CVCV, CVCVC, or CVCVCV format, while the pronunciation lengths of the two pseudo-words are the same in each question. Unlike the phoneme experiment, pronunciation of two pseudo-words in each question is not necessarily similar

(e.g., /'louf/ and /'gid/). Therefore, this experiment can measure the performance of the proposed models for pronunciation length with more arbitrary pseudo-words compared to the phoneme experiment.

4.1 Preparing Pseudo-words and Questions

In each trial of the experiments, we show participants a pair of words that do not exist in the English vocabulary (pseudo-words). For this reason, we create 400 pseudo-words to prepare 200 questions for each of the experiments.

4.1.1 Pseudo-words for the Phoneme Experiment

The phoneme experiment only uses pseudo-words in CVCV format. This experiment measures the performance towards types of phonemes. Additionally, each question (a pair of pseudo-words A and B) should be in either of the following two formats:

- Vs differ ($C_1V_A C_2V_A$ and $C_1V_B C_2V_B$; e.g., /'fu.u/ and /'fi.i/)
- Cs differ ($C_A V_1 C_A V_2$ and $C_B V_1 C_B V_2$; e.g., /'bubaʊ/ and /'wuwaʊ/)

Besides, in this experiment, each pseudo-word must satisfy all of the following four conditions:

1. The first syllable is always stressed
2. The pronunciation is not the same as the pronunciation of existing English words
3. The pronunciation is not the same as Bouba, Maluma, Kiki, or Takete
4. The pronunciation has high wordiness [13] in English

4.1.2 Pseudo-words for the Length Experiment

The length experiment uses pseudo-words either in CVC, CVCV, CVCVC, or CVCVCV format. This experiment tests the influence of pronunciation length. More specifically, each question (a pair of pseudo-words A and B) should be in either of the following four formats:

- Comparison of CVC words
($C_{A_1} V_{A_1} C_{A_2}$ and $C_{B_1} V_{B_1} C_{B_2}$; e.g., /'louf/ and /'gid/)
- Comparison of CVCV words
($C_{A_1} V_{A_1} C_{A_2} V_{A_2}$ and $C_{B_1} V_{B_1} C_{B_2} V_{B_2}$; e.g., /'zʊgəʊ/ and /'muθə/)
- Comparison of CVCVC words
($C_{A_1} V_{A_1} C_{A_2} V_{A_2} C_{A_3}$ and $C_{B_1} V_{B_1} C_{B_2} V_{B_2} C_{B_3}$; e.g., /'mifeil/ and /'θoʊfʊθ/)
- Comparison of CVCVCV words
($C_{A_1} V_{A_1} C_{A_2} V_{A_2} C_{A_3} V_{A_3}$ and $C_{B_1} V_{B_1} C_{B_2} V_{B_2} C_{B_3} V_{B_3}$; e.g., /wɪ'kupə/ and /'vʌbəzi/)

Besides, in this experiment, each pseudo-word must satisfy all of the following two conditions:

1. The pronunciation is not the same as Bouba, Maluma, Kiki, or Takete
2. The pronunciation has high wordiness [13] in English

Table 1: All phonemes used to create pseudo-words.

Consonants	/p/, /t/, /k/, /f/, /θ/, /s/, /ʃ/, /b/, /d/, /g/, /v/, /ð/, /z/, /ʒ/, /m/, /n/, /ɹ/, /l/, /w/, /j/
Vowels	/ei/, /i/, /ai/, /oʊ/, /u/, /aʊ/, /ɔɪ/, /ɑ/, /æ/, /ɛ/, /ɪ/, /ʊ/, /ʌ/, /ə/

In contrast to the phoneme experiment, here we omit the condition that “the pronunciation is not the same as the pronunciation of existing English words.” This is because the condition is too strict to ensure the phonetic variety of the created words in this experiment, especially in CVC format. In CVC (C_1VC_2) format, for example, the case where C_1 is /k/ and C_2 is /t/ is quite popular in English (e.g., “Cat”:/kæt/, “Kit”:/kɪt/, “Coat”:/koʊt/), while the case where C_1 is /θ/ and C_2 is /ʒ/ is quite rare. Therefore, including the condition would result in a lot of /θ/+V+/ʒ/ words and few /k/+V+/t/ words, which can work as a negative bias throughout the evaluation.

4.1.3 Question Generation

For each of the experiments, we first create all possible pseudo-words that satisfy all of the conditions above, and then randomly select 400 words from them to prepare 200 questions in the aforementioned format. These pseudo-words are then converted into audio sounds to be presented to participants.

We use the software PronunDict² to check the overlap of pronunciation with existing words in the phoneme experiment. Also, we use Speech Application Programming Interface (SAPI)³ to convert pronunciation written with International Phonetic Alphabet (IPA) into spoken sounds. Here, Zira (American English, Female) is selected as the speaker, and the speed of her speech is fixed by setting the speaking rate as -2 . Besides, to ensure high wordiness of the pseudo-words, we refer to American English phonology from J.C. Wells [29] for phoneme selection and pseudo-word format creation. All phonemes used in the experiments are shown in Table 1. Note that several phonemes having problems in converting to audio sounds are excluded from the experiments such as /h/ and /ŋ/⁴.

4.2 Experimental Settings

Both experiments are conducted targeting English native speakers in a crowd-sourcing fashion via Amazon Mechanical Turk⁵.

In both of the experiments, participants are first asked to listen to two pseudo-words as audio files. Then, they are asked to respond to which pronunciation sounds pointier, with available responses being: “Audio 1,” “Audio 2,” and “Identical.” The actual user interface displayed to participants is shown in Fig. 7. We collect five such answers for each question and determine a

²<https://sourceforge.net/projects/pronundict/>

³[https://docs.microsoft.com/en-us/previous-versions/windows/desktop/ee125663\(v=vs.85\)](https://docs.microsoft.com/en-us/previous-versions/windows/desktop/ee125663(v=vs.85))

⁴The 400 pseudo-words used in each of the experiments are publicly available at <https://www.cs.is.i.nagoya-u.ac.jp/files/misc/matsuhirac/phoneticpointiness/>.

⁵<https://www.mturk.com/>

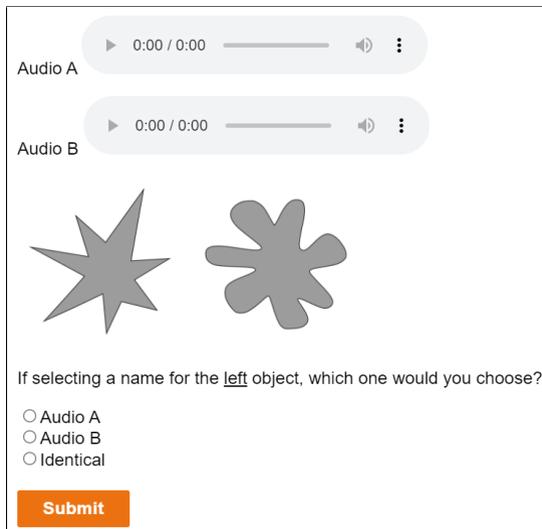


Fig. 7: User interface shown to participants in the crowd-sourcing experiments.

ground-truth answer for each question by taking a majority vote of the five answers.

Besides, we categorize each question by the degree of consensus into four categories: *Perfect* (all agree), *Strong* (four out of five agree), *Weak* (three out of five agree), and *Others* (less than three out of five agree) categories. This categorization reflects the difficulty of the questions from high consensus to low consensus between participants.

For the evaluation, we use questions only in the *Perfect*, *Strong*, and *Weak* categories (we name the union of the three categories as *All* category) and do not use questions in the *Others* category, in which participants' answers did not agree well (14 and 17 questions in each experiment, respectively). We excluded these less-agreed questions for evaluation, since even humans would not be able to agree on a correct answer. In both experiments, we measure accuracy as an evaluation metric. In the phoneme experiment, however, we found that there is a gap between the numbers of questions in which only vowels differ and only consonants differ. Thus, instead of accuracy, we calculate MacroAcc, a macro average accuracy of two accuracies Acc_V and Acc_C . The former, Acc_V , is an accuracy of questions in which only vowels differ and the latter, Acc_C , is an accuracy of questions in which only consonants differ. The metric MacroAcc is calculated as

$$\text{MacroAcc} = \frac{Acc_V + Acc_C}{2}. \quad (8)$$

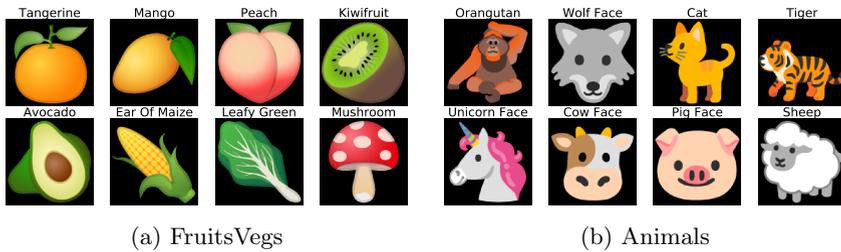


Fig. 8: Example images and tags in emoji subdatasets

4.3 Training Data for Image Generator of Referential Model

The referential model requires many pairs of a pronunciation and a silhouette image of existing words. As such data, we use Google Emoji dataset⁶, which consists of more than 1,000 words with emoji images with a resolution of 136×136 pixels. Here, we use tags attached to each emoji to determine the name of the object displayed in the emoji image.

This dataset provides emojis not only for general objects but also for facial expressions e.g., “Smiling Face” and “Angry Face.” Although the attached tags are different, all such emojis related to facial expressions have the same circular-shaped face silhouette. This means that the pronunciation of the tags might not necessarily relate to the silhouette shapes of the objects. Thus, in this paper, we filter out several words from the whole vocabulary of the dataset and make three sub-datasets: FruitsVegs, Animals, and FruitsVegs+Animals.

FruitsVegs consists of 34 emojis/words that illustrate either fruits or vegetables, and Animals consists of 111 emojis/words that illustrate animals. Example images and tags of these categories are shown in Figs. 8a and 8b, respectively. FruitsVegs+Animals consists of 145 emojis/words from those two sub-datasets.

Lastly, we manually modify several tags of emojis in these sub-datasets to be more popular and consistent ones. For example, we modify “Kiwifruit” to “Kiwi,” which the fruit is casually called, and “Wolf Face” to “Wolf” so that the entire name represents only the name of the illustrated animal⁷.

Figure 8 shows examples of pairs of an emoji and a raw tag in each sub-dataset. For training the image generator, we use pronunciation data converted from modified tags as well as silhouette images extracted from raw emoji images. We use the alpha channel of each image to extract its silhouette.

⁶<https://github.com/iamcal/emoji-data/>

⁷All the modifications made to the raw tags are also available at <https://www.cs.i.nagoya-u.ac.jp/files/misc/matsuhirac/phoneticpointiness/>.

Table 2: Combinations of W_r and W_p used in the experiments

Abbreviation	W_r	W_p
BK	{Bouba}	{Kiki}
MT	{Maluma}	{Takete}
BMKT	{Bouba, Maluma}	{Kiki, Takete}

4.4 Pronunciation Sets Used in Expressive Model

As the candidate elements of pronunciation sets W_r and W_p of the proposed expressive model, we adopt well-known pseudo-words Bouba and Maluma for W_r , and Kiki and Takete for W_p . In the experiments, we test three combinations of W_r and W_p , which are shown in Table 2.

4.5 Implementation Details

We perform data augmentation for emoji silhouette images used to train the DF-GAN of the proposed referential model. In detail, by rotating images by 10 degrees, we augment each image into 36 rotated images. In order to decide the optimal training epoch of the DF-GAN, we calculate Fréchet Inception Distance (FID) [30], which is a major metric to measure the performance of image generation models. Eventually, we use the epoch that gives the lowest, i.e., the best FID. The resolution of generated images is set as 256×256 pixels.

As the data to train the Word2vec and the Auto-Encoder of the proposed expressive model, we use Spell Checker Oriented Word Lists (SCOWL)⁸, an English wordlist that contains 44,561 words. We set the number of dimensions of Word2vec word embeddings as 200, and of the phonetic space generated by the Auto-Encoder as 256. Word2vec is trained using Skip-gram. The Encoder of the Auto-Encoder including the Word2vec is also adopted as the Text-Encoder of the DF-GAN used in the referential model. Note that we also perform normalization to the embedding obtained by the Encoder before feeding them into the DF-GAN, so that the scale matches with that of the original Text-Encoder of the DF-GAN.

For converting English words into pronunciation written with International Phonetic Alphabet (IPA), we use a Python package `eng-to-ipa`⁹. With regard to this package, we found that the conversion does not support the phoneme $/\Lambda/$, so the pronunciation of words that contains $/\Lambda/$ is always described with $/\text{ə}/$ instead (e.g., “Hut” is converted into $/\text{h}\text{ət}/$ instead of $/\text{h}\Lambda\text{t}/$). Considering this, we replace all $/\Lambda/$ in pseudo-words with $/\text{ə}/$ just before inputting them into the proposed models. Due to this additional conversion, for example, a pseudo-word $/\text{ʌ}\text{n}\Lambda\text{u}\text{v}/$ is converted into $/\text{ʌ}\text{n}\text{ə}\text{u}\text{v}/$ before it is input to each model.

According to the phonology based on J.C. Wells [29], we set the actual pronunciations of Bouba, Maluma, Kiki, and Takete as $/\text{b}\text{u}\text{b}\text{ə}/$, $/\text{m}\Lambda\text{l}\text{u}\text{m}\Lambda/$, $/\text{k}\text{i}\text{k}\text{i}/$, $/\text{t}\Lambda\text{k}\text{e}\text{t}\text{ə}/$, respectively. The distribution of these pronunciations on

⁸<http://wordlist.aspell.net/>

⁹<https://pypi.org/project/eng-to-ipa/>

Table 3: Results of the two experiments. Each combined model uses the corresponding referential model and the asterisked expressive model in each experiment.

Model	Vocabulary	Accuracy \uparrow (Number of questions)			
		Phoneme (186)		Length (183)	
Referential	FruitsVegs	0.409		0.459	
	Animals	0.456		0.470	
	FruitsVegs+Animals	0.580		0.497	
Expressive	BK	0.591		0.514	
	MT	0.604		0.541	
	BMKT	0.652*		0.541*	
Combined	FruitsVegs	0.665	($a = 0.144$)	0.541	($a = 0.000$)
	Animals	0.657	($a = 0.038$)	0.557	($a = 0.266$)
	FruitsVegs+Animals	0.691	($a = 0.280$)	0.585	($a = 0.692$)

expressive model and the combined model perform always better than random choice. The referential model, on the other hand, appears to be struggling in almost all contexts. In addition, in both experiments, the expressive model always performs better than the referential model. These suggest that human perception towards phonetic pointiness can be explained better by the expressive sense (human intuition) than the referential sense (reference to existing words).

Regarding the referential model, we can see that the vocabulary used in it affects the performance very much. The vocabulary size of each sub-dataset is 34 (FruitsVegs), 111 (Animals), and 145 (FruitsVegs+Animals). Considering this, in both experiments, it appears that a larger vocabulary improves the performance. Because humans have a much larger vocabulary than all of the sub-datasets used in the experiments, these results indicate that the referential model may approximate human perception better if the vocabulary increases.

Regarding the expressive model, we can see that the choice of W_r and W_p affects the performance. In both experiments, we can observe that only using Maluma and Takete performs better than only using Bouba and Kiki, and also that using all of Bouba, Maluma, Kiki, and Takete makes the performance much better. This means that, for English speakers, the pair of Maluma and Takete is a better example of pointy/round association than Bouba and Kiki, in terms of concordance with human perception. Also, the result suggests the possibility of the existence of more effective pronunciation pairs that should be included in W_r and W_p . Based on the fact that the expressive model fits with human perception well, such a more effective pair would increase the accuracy if integrated into the expressive model. Hence, by observing the behavior of this expressive model testing with different combinations of W_r and W_p , it would be interesting to discover which pronunciation pair is the most effective.

Lastly, as we have expected, the combined model performs much better than the original two models, giving the highest MacroAcc of 0.691 for the phoneme experiment and the highest accuracy of 0.585 for the length experiment. By observing its weight a , we can grasp how much the referential and

Table 4: Detailed results of the phoneme experiment arranged by categories and types of questions.

Category	Type (Number of questions)	Accuracy and MacroAcc \uparrow		
		Referential (FruitsVegs+Animals)	Expressive (BMKT)	Combined ($a = 0.280$)
All	Vowels+Consonants (186)	0.580	0.652	0.691
	Vowels (106)	0.585	0.642	0.670
	Consonants (80)	0.575	0.663	0.713
Perfect	Vowels+Consonants (24)	0.500	0.879	0.879
	Vowel (14)	0.500	0.857	0.857
	Consonants (10)	0.500	0.900	0.900
Strong	Vowels+Consonants (57)	0.614	0.685	0.736
	Vowel (28)	0.607	0.714	0.714
	Consonants (29)	0.621	0.655	0.759
Weak	Vowels+Consonants (105)	0.577	0.586	0.622
	Vowel (64)	0.594	0.563	0.609
	Consonants (41)	0.561	0.610	0.634

the expressive features, respectively, are important for each experiment. In the phoneme experiment, the optimal weight a always ranges near 0, representing that the expressive sense is always more important than the referential sense. Interestingly, in the length experiment, the optimal weight a of the best model that uses FruitsVegs+Animals is 0.692. This is a contrasting result from the phoneme experiment, implying the larger importance of the referential information. The pseudo-words used in the length experiment have a wider phonetic variety than those used in the phoneme experiment, where the format of pseudo-words is strictly restricted to CVCV. From this point of view, we assume that the similarity of the phonetic distribution between pseudo-words and existing words may have become closer in the length experiment, which may have resulted in such a high value of the optimal weight a .

4.6.2 Results by Question Type

Next, we discuss the types of questions. The detailed results of the phoneme experiment are shown in Table 4. The table shows MacroAccs given by the best models of Table 3 arranged by categories (*All*, *Perfect*, *Strong*, and *Weak*; details have been described in Sec. 4.2) and types of questions (within two pseudo-words, whether vowels differ or consonants differ).

First, in all models, there seems to be no remarkable difference between strengths towards questions in which vowels differ or consonants differ. On the other hand, for categories, we can see that the referential model has strengths mainly in the *Strong* and *Weak* categories, while the expressive model is significantly strong in the *Perfect* category. The *Perfect* category should contain obvious questions similar to “Bouba vs. Kiki,” which can be a reason for such high performance of the expressive model, since it has prior knowledge of the Bouba-Kiki effect. The referential model, on the other hand, does not have such

a prior, and just tries to guess answers only from the knowledge of the names and shapes of fruits, vegetables, and animals. Considering this, the results of the referential model suggest that the pronunciations of such existing English words do certainly represent their shape information, since the model correctly guessed many questions in the *Strong* and *Weak* categories. However, since it cannot perform well on the *Perfect* category, the information embedded in existing words is not obvious enough to explain the Bouba-Kiki effect by itself. This can also be one of the reasons why the combined model achieved better performance than the other two models.

The table also shows that the combined model performs the best in every category, giving the highest MacroAcc of 0.691, 0.879, 0.736, and 0.622 in each category. Especially in the *Strong* and *Weak* categories, we can observe a great improvement of performance from the rest of the two models. The possible reason is already mentioned in the last paragraph. Focusing on the performance itself, supposing that humans answer the same questions, the expectation of their accuracy should theoretically be 1.000, 0.800, and 0.600 for the *Perfect*, *Strong*, and *Weak* categories, respectively. Comparing these accuracies with the combined model indicates that it performs very well. These results suggest the applicability of the combined model to other tasks such as estimation and analysis of the phonetic pointiness, which will be discussed further in Section 5.

Next, the detailed results of the length experiment are shown in Table 5. The table shows accuracies given by the best models of Table 3 arranged by categories (*Perfect*, *Strong*, and *Weak*) and types of questions (questions in the format of CVC, CVCV, CVCVC, and CVCVCV).

In terms of categories, we can see a similar tendency to Table 4. Besides, regarding pronunciation length, we can also confirm that the referential model has strengths in the CVC format questions, whereas the expressive model has weaknesses. This suggests that the pronunciation of existing words in the CVC format represents their shape information more than words in other formats. For example, a “Pig” has a relatively round shape and its pronunciation /'pɪg/ might also evoke a round impression, while a “Fish” has a pointier shape and its pronunciation /'fɪʃ/ also evokes a pointier impression. We believe that our model successfully handled such implicit features of the natural language.

Moreover, we can see that the expressive model has strength in the CVCVCV format questions. This is probably because we used prior knowledge of Maluma and Takete when constructing the expressive model. These two words are known as contrasting examples of two CVCVCV words with which the Bouba-Kiki effect can be observed. This must have worked effectively for the model to guess human answers correctly.

Finally, also in the length experiment, the combined model achieves higher accuracies than the other two models throughout almost all categories and types of questions. However, it is still struggling to achieve high performance, as the question setting is much more complicated than the phoneme experiment. When the detail is focused, the model performs especially badly on the

Table 5: Detailed results of the length experiment arranged by categories and types of questions.

Category	Type (Number of questions)	Accuracy \uparrow			
		Referential (FruitsVegs+Animals)	Expressive (BMKT)	Combined ($a = 0.692$)	
All	All questions	(183)	0.497	0.541	0.585
	CVC	(46)	0.652	0.478	0.609
	CVCV	(48)	0.500	0.521	0.604
	CVCVC	(43)	0.465	0.558	0.628
	CVCVCV	(46)	0.370	0.609	0.500
Perfect	All questions	(15)	0.267	0.600	0.467
	CVC	(3)	0.667	0.333	0.667
	CVCV	(3)	0.333	1.000	0.667
	CVCVC	(5)	0.000	0.600	0.200
	CVCVCV	(4)	0.250	0.500	0.500
Strong	All questions	(50)	0.540	0.520	0.640
	CVC	(21)	0.524	0.476	0.524
	CVCV	(7)	0.857	0.571	0.857
	CVCVC	(12)	0.417	0.500	0.667
	CVCVCV	(10)	0.500	0.600	0.700
Weak	All questions	(118)	0.508	0.542	0.576
	CVC	(22)	0.773	0.500	0.682
	CVCV	(38)	0.447	0.474	0.553
	CVCVC	(26)	0.577	0.577	0.692
	CVCVCV	(32)	0.344	0.625	0.438

CVCVCV format questions, giving a neutral accuracy of 0.500 in the *All* category. This can be the weakness of the current combined model. This weakness is likely to be caused by its combination scheme. Therefore, implementing a more sophisticated combination scheme would effectively solve this problem. For example, because the current model fixes its weight a for any kind of input pronunciation, adjusting a dynamically depending on the input might improve the performance.

5 Applications

This section discusses the applicability of the proposed models by focusing on two possible use-cases. Section 5.1 explores the usefulness of the models when naming commercial products. Section 5.2 explores a deeper aspect of the Bouba-Kiki effect by observing the behavior of the models towards a set of certain input pronunciations.

5.1 Product Naming

One of the applications of this research would be assisting people to give names to commercial products. The main use-case of the proposed models would be; given many candidate names, the models can help people choose one or several names from the candidates that best match the characteristics of the product.

Table 6: Examples of calculated phonetic pointiness scores by the proposed models for synonym words.

Word	Pronunciation	Referential \hat{S}_{Ref} (FruitsVegs+Animals)	Expressive \hat{S}_{Exp} (BMKT)	Combined S_{Cmb} ($a = 0.280$)
Sharp	/ʃɑ:p/	-1.087	-0.629	-0.757
Thorny	/θɔ:ni/	0.342	0.698	0.598
Spiky	/ˈspaɪki/	-1.685	1.544	0.640
Pointy	/ˈpɔɪnti/	1.893	1.083	1.310

In this section, in order to show an example of such use-cases, we calculate the phonetic pointiness scores by the proposed models towards certain synonym adjectives. The calculated scores are shown in Table 6. All these four adjectives have a meaning related to the pointiness of objects. Note that, to demonstrate the usefulness of the proposed models for unseen words, we chose only words that are not used to train the image generator of the referential model.

In the table, among the four adjectives related to pointiness, “Pointy” shows the highest phonetic pointiness according to the referential model. In contrast, the scores calculated by the expressive model suggest that “Spiky” sounds the pointiest. In addition, the combined model integrates the two models and concludes that “Pointy” would sound pointier than “Spiky” to humans.

These results imply that pure human intuition perceives “Spiky” the pointiest, while “Pointy” is more likely to evoke existing *namely* pointy words than “Spiky” (e.g., **pineapple** (/ˈpaɪnæpəl/)). Also, since the combined model is optimized to estimate the reaction of an average human, the optimal combined model tells that an average human would perceive “Pointy” pointier than “Spiky.”

Imagine when we need to name a product whose characteristic is its spiky shape. From the results above, we can see that naming “Spiky” to the product would best match its characteristic in terms of *human intuition*. In contrast, when we want to consider *the referential aspect* of sound symbolism, it would be better to choose “Pointy” rather than “Spiky” for the product name as both the referential model and the optimal combined model suggest. In any case, it can also be deduced that naming the product “Sharp” seems inappropriate for this context.

By using the proposed models, we can estimate such human reactions without market research, or even when we are not familiar with the target language. Since these methods can also be applied to arbitrary pronunciations including words that do not exist, the proposed models would be a helpful tool for people to name commercial products. In addition, according to the intention of the naming, people have a choice of which model to use for assessing the candidate names. For example, the expressive model is useful if people want to know how pointy the names intuitively sound, and the referential model and the combined model are useful if they want to consider the referential symbolism and assess in terms of more general human perception.

Table 7: Five most representative pseudo-words in each distribution group of phonetic pointiness.

Rank	← Low (Round)			High (Pointy) →	
	0%–20%	20%–40%	40%–60%	60%–80%	80%–100%
1	/'buba/	/'ʒʊjʊʊ/	/'ðʊʊvɔɪ/	/'dɔɪdeɪ/	/'kiker/
2	/'juba/	/'jʊjʊʊ/	/'ðəʊvɔɪ/	/'dɔɪdai/	/'kekeɪ/
3	/'bæba/	/'ʒʊzʊʊ/	/'ɪʊʊvɔɪ/	/'deɪdeɪ/	/'kikar/
4	/'jæba/	/'jʊzʊʊ/	/'ɪəʊvɔɪ/	/'θɔɪdeɪ/	/'kɔɪkeɪ/
5	/'baba/	/'ʒəʊjʊʊ/	/'ðʊʊgɔɪ/	/'deɪdai/	/'kekai/

5.2 Exploring Pronunciation between Bouba and Kiki

As the Bouba-Kiki effect states, the pronunciation of Bouba evokes a round impression and the pronunciation of Kiki evokes a pointy impression. What about other pronunciations? Psycholinguistics has yet to provide a definite answer to this question. In this section, we attempt to explore what the distribution of pronunciation between Bouba and Kiki is like by observing the behavior of the proposed combined model for pseudo-words in the CVCV format.

First, we generate all possible $C_1V_1C_2V_2$ pseudo-words according to English phonology based on J.C. Wells [29] as performed in Section 4.1. Next, we calculate the phonetic pointiness scores for all these pseudo-words with the combined model (with the optimal $a = 0.280$), and sort them by their phonetic pointiness scores. Then, in order to observe the tendency of the distribution rather than the absolute rankings of the pseudo-words, we classify all pseudo-words into five groups by their phonetic pointiness scores: 0%–20%, 20%–40%, 40%–60%, 60%–80%, and 80%–100% of the data distribution. For example, if the phonetic pointiness score of a word is equal to or greater than the first quintile (20%) but less than the second quintile (40%) of the distribution, the word is classified into the 20%–40% group and the word is regarded as having a relatively low phonetic pointiness. Finally, by analyzing these classified pseudo-words, we search for the most representative pseudo-words in each distribution group. In detail, when $f_{C_1}(X)$, $f_{V_1}(X)$, $f_{C_2}(X)$, and $f_{V_2}(X)$ represent discrete probability distributions of a phoneme variable X in each position of a pseudo-word $C_1V_1C_2V_2$, we calculate a likelihood $f_{C_1}(X = c_1)f_{V_1}(X = v_1)f_{C_2}(X = c_2)f_{V_2}(X = v_2)$ for all possible pseudo-words $c_1v_1c_2v_2$. Then, we search for the top five pseudo-words that give the maximum likelihood. Note that the occurrence of phonemes in C_1 , V_1 , C_2 , and V_2 positions are supposed to be independent throughout the calculation.

The five most representative pseudo-words in each group are shown in Table 7. First, as we have expected, Bouba-like pseudo-words (e.g., /'juba/) appear in group 0%–20% and Kiki-like pseudo-words (e.g., /'kiker/) appear in group 80%–100%, which means that such Bouba-like words tend to have the lowest scores, and Kiki-like words tend to have the highest scores. Next, regarding vowels, we can see that back vowels and rounded vowels (/u/, /ʊ/, /o/, /ɔ/, /ɑ/) appear frequently in low phonetic pointiness groups, while close

front unrounded vowels (/i/, /ɪ/, /e/, /ɛ/) appear frequently in high phonetic pointiness groups. Regarding consonants, sonorants (/j/, /ɹ/) and voiced consonants (/b/, /ɓ/, /z/, /ð/, /v/, /d/, /g/) appear frequently in low phonetic pointiness groups, while voiceless consonants (/k/, /θ/) appear frequently in high phonetic pointiness groups. These results all match the existing findings of the Bouba-Kiki effect [15, 17], thus we can confirm that the proposed combined model performs quite well in terms of the capability of handling phonemes, too.

According to this table, pseudo-words having both round phonemes (i.e., phonemes that are often associated with round shapes) and pointy phonemes tend to be in the 40%–60% group, which is exactly the midpoint between Bouba and Kiki (e.g., /ðouɹɔ/ has a round phoneme /ou/ and a pointy phoneme /ɹɔ/). This is reasonable since such words have both features that can be perceived as either round or pointy.

Following this, we observed that the optimal combined model can imitate the human reaction towards the Bouba-Kiki effect quite well. The analysis of the combined model succeeded in revealing the tendency of the pronunciation and phoneme occurrences over the model outputs. Therefore, it would also be interesting to compare the tendency among phonetic impressions other than the phonetic pointiness, such as how hard/soft or how big/small a pronunciation is perceived.

6 Conclusion

In this paper, we introduced the novel concept of “phonetic pointiness” as how pointy a shape humans are most likely to associate with a given pronunciation. Based on psycholinguistic findings arguing types of sound symbolism, we proposed three models, a referential model, an expressive model, and a combined model of the previous two, to calculate the phonetic pointiness for an arbitrary pronunciation. This was the first attempt to model the phonetic pointiness. For this, we employed computational and data-driven approaches incorporating referential symbolism and expressive symbolism.

We evaluated the proposed models through subjective evaluation by estimating human answers towards which of the two pronunciations sounds pointier. The results showed a promising correlation between human perception and the phonetic pointiness calculated by the proposed models. The combined model improved the performance compared to the sole models, indicating that human perception can be modeled more accurately by combining the two types of sound symbolisms.

We showed several use-cases of the proposed models, which showcased the applicability of this research not only in informatics and engineering but also in marketing and psycholinguistics fields. By further utilizing and analyzing the proposed models, we believe that much more applications can be discovered, such as analyzing cultural differences among language speakers and revealing the word origins of natural languages.

In actual applications, users have a choice of which model to use depending on their use-cases, since there are models for two different symbolisms and the combined model. This is hard to achieve if the model is constructed via training with human annotations, i.e., supervised learning. The reason for this is that it requires task-specific human annotations for each symbolism, which could be hard to collect because of the ambiguity of the boundary of the two symbolisms. Therefore, this usage would be a great advantage against the current trend of training a model under the supervision of manually annotated data.

Nevertheless, we recognize that the proposed models have some limitations. First of all, the performance, especially of the current referential model, might not be stable enough for some applications. It is mainly influenced by three factors: (1) the performance of the image generator, (2) the performance of the visual pointiness calculator, and (3) the data used to train the image generator. For the first and second factors, better and more accurate methods should be adopted, or proposed. For the third one, other data should be considered for the training, since the vocabulary of the emoji dataset used in this paper is quite limited. The restriction that the data must satisfy is that each name of an object is attached to its silhouette images. In addition, it would be better if the pronunciation of the names relates very much to the silhouette shapes of the objects.

For the expressive model, one possible point for improvement is the phonetic space construction. Currently, Word2vec is used to obtain the phonetic/phonological relationship among phonemes. Although embeddings of Word2vec can capture such relationships to some extent [27], there should be better approaches such as directly integrating phonetic/phonological priors of phonemes such as IPA chart [7] in place of the Word2vec embeddings.

Another limitation we found for the combined model, is the combination scheme. The current scheme is simple and effective, but it would be better if the weight a were automatically determined based on psycholinguistic findings. Moreover, because the current combined model fixes its weight a for any kind of input pronunciation, adopting a dynamic weighting scheme when combining scores would add more flexibility to the output of the model.

Considering these limitations, for future work, our next step is to extend the framework of the proposed models to be able to estimate phonetic impressions other than pointiness, such as how hard/soft or how big/small a pronunciation is perceived. We believe that the framework of the referential model, combined with other vision and language models, has the potential to achieve this. Since the use of silhouette images of objects is not suitable for expressing such visual impressions evoked by pronunciation, the color and texture information of the objects should also be taken into account. Besides, by comparing the behavior of models constructed in different languages, we could reveal the difference in people's reactions towards phonetic impressions in different languages and different cultural backgrounds. Furthermore, among different phonetic impressions, it would also be interesting to compare the behavior of models, as made in Section 5.2.

Statements and Declarations

This work was partly supported by Microsoft Research CORE16 program, JSPS Grant-in-aid for Scientific Research (22H03612), and Nagoya University Interdisciplinary Frontier Fellowship supported by Japan Science and Technology Agency and Nagoya University.

This work is a fruit of a joint research project between Nagoya University and the National Institute of Informatics.

Compliance with Ethical Standards

This research was approved by the Institutional Review Board at the Graduate School of Informatics, Nagoya University. Informed consent was obtained from each participant.

Data Availability Statements

The datasets generated during and/or analysed during the current study are available at <https://www.cs.is.i.nagoya-u.ac.jp/files/misc/matsuhirac/phoneticpointiness/>.

References

- [1] Hinton, L., Nichols, J., Ohala, J.J.: Sound Symbolism. Cambridge University Press, Cambridge, England, UK (1995). <https://doi.org/10.1017/CBO9780511751806>
- [2] Köhler, W.: Gestalt Psychology. H. Liveright, New York, NY, USA (1929)
- [3] Ramachandran, V.S., Hubbard, E.M.: Synaesthesia —A window into perception, thought and language. *J. Conscious. Stud.* **8**(12), 3–34 (2001)
- [4] Papantoniou, K., Konstantopoulos, S.: Unravelling names of fictional characters. In: Proc. 54th Annual Meet. Assoc. Comput. Linguist., vol. 1, pp. 2154–2163 (2016). <https://doi.org/10.18653/v1/P16-1203>
- [5] Sabbatino, V., Troiano, E., Schweitzer, A., Klinger, R.: “splink” is happy and “phrouth” is scary: Emotion intensity analysis for nonsense words. In: Proc. 12th Workshop Comput. Approaches to Subj. Sentiment Soc. Media Anal., pp. 37–50 (2022). <https://doi.org/10.18653/v1/2022.wassa-1.4>
- [6] Sapir, E.: A study in phonetic symbolism. *J. Exp. Psychol.* **12**(3), 225–239 (1929). <https://doi.org/10.1037/h0070931>
- [7] International Phonetic Association: Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge University Press, Cambridge, England, UK (1999)

- [8] Maurer, D., Pathman, T., Mondloch, C.: The shape of Boubas: Sound-shape correspondences in toddlers and adults. *Dev. Sci.* **9**(3), 316–322 (2006). <https://doi.org/10.1111/j.1467-7687.2006.00495.x>
- [9] Ozturk, O., Krehm, M., Vouloumanos, A.: Sound symbolism in infancy: Evidence for sound-shape cross-modal correspondences in 4-month-olds. *J. Exp. Child Psychol.* **114**(2), 173–186 (2013). <https://doi.org/10.1016/j.jecp.2012.05.004>
- [10] Bremner, A.J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K.J., Spence, C.: ‘Bouba’ and ‘Kiki’ in Namibia? A remote culture make similar shape-sound matches, but different shape-taste matches to Westerners. *Cognition* **126**(2), 165–172 (2013). <https://doi.org/10.1016/j.cognition.2012.09.007>
- [11] Chen, Y.C., Huang, P.C., Woods, A., Spence, C.: When ‘Bouba’ equals ‘Kiki’: Cultural commonalities and cultural differences in sound-shape correspondences. *Sci. Rep.* **6**, 26681–1266819 (2016). <https://doi.org/10.1038/srep26681>
- [12] Ćwiek, A., Fuchs, S., Draxler, C., Asu, E.L., Dediu, D., Hiovain, K., Kawahara, S., Koutalidis, S., Krifka, M., Lippus, P., Lupyán, G., Oh, G., Paul, J., Petrone, C., Ridouane, R., Reiter, S., Schümchen, N., Szalontai, Á., Ünal-Logacev, Ö., Winter, B.: The Bouba/Kiki effect is robust across cultures and writing systems. *Philos. Trans. Royal Soc. B Biol. Sci.* **377** (2021). <https://doi.org/10.1098/rstb.2020.0390>
- [13] Styles, S.J., Gawne, L.: When does Maluma/Takete fail? Two key failures and a meta-analysis suggest that phonology and phonotactics matter. *i-Perception* **8**(4), 17p. (2017). <https://doi.org/10.1177/2041669517724807>
- [14] Rogers, S.K., Ross, A.S.: A cross-cultural test of the Maluma-Takete phenomenon. *Perception* **4**(1), 105–106 (1975). <https://doi.org/10.1068/p040105>
- [15] McCormick, K., Kim, J.Y., List, S., Nygaard, L.: Sound to meaning mappings in the Bouba-Kiki effect. In: *Proc. 37th Annual Meet. Cogn. Sci. Soc.*, pp. 1565–1570 (2015)
- [16] Fort, M., Martin, A., Peperkamp, S.: Consonants are more important than vowels in the Bouba-Kiki effect. *Lang. Speech* **58**(2), 247–266 (2015). <https://doi.org/10.1177/0023830914534951>
- [17] Monaghan, P., Mattock, K., Walker, P.: The role of sound symbolism in language learning. *J. Exp. Psychol. Learn. Mem. Cogn.* **38**(5), 1152–1164 (2012). <https://doi.org/10.1037/a0027747>

- [18] Knoeferle, K., Li, J., Maggioni, E., Spence, C.: What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings. *Sci. Rep.* **7**(1), 5562_1–5562_11 (2017). <https://doi.org/10.1038/s41598-017-05965-y>
- [19] Gillet, O., Richard, G.: Drum loops retrieval from spoken queries. *J. Intell. Inf. Syst.* **24**(2), 159–177 (2005). <https://doi.org/10.1007/s10844-005-0321-9>
- [20] Sundaram, S., Narayanan, S.: Classification of sound clips by two schemes: Using onomatopoeia and semantic labels. In: *Proc. 2008 IEEE Int. Conf. Multimed. Expo*, pp. 1341–1344 (2008). <https://doi.org/10.1109/ICME.2008.4607691>
- [21] Fukusato, T., Morishima, S.: Automatic depiction of onomatopoeia in animation considering physical phenomena. In: *Proc. 7th Int. Conf. Motion Games*, pp. 161–169 (2014). <https://doi.org/10.1145/2668064.2668096>
- [22] Tao, M., Tang, H., Wu, S., Sebe, N., Wu, F., Jing, X.Y., Bao, B.: DF-GAN: Deep Fusion Generative Adversarial Networks for text-to-image synthesis. *Comput. Res. Reposit.*, arXiv preprint, arXiv:2008.05865 (2020). <https://doi.org/10.48550/arXiv.2008.05865>
- [23] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: *Adv. Neural Inf. Process. Syst.*, vol. 27, 9p. (2014)
- [24] Matsuhira, C., Kastner, M.A., Ide, I., Kawanishi, Y., Hirayama, T., Doman, K., Deguchi, D., Murase, H.: Pointedness of an image: Measuring how pointy an image is perceived. In: *HCI Int. Conf. 2021 —Posters, Procs., Part I, Commun. Comput. Inf. Sci.*, vol. 1419, pp. 137–144 (2021). https://doi.org/10.1007/978-3-030-78635-9_20
- [25] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. *Comput. Res. Reposit.*, arXiv preprint, arXiv:1301.3781 (2013). <https://doi.org/10.48550/arXiv.1301.3781>
- [26] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Adv. Neural Inf. Process. Syst.*, vol. 26, pp. 3111–3119 (2013)
- [27] Kolachina, S., Magyar, L.: What do phone embeddings learn about Phonology? In: *Proc. 16th Workshop Comput. Res. Phonetics, Phonol., Morphol.*, pp. 160–169 (2019). <https://doi.org/10.18653/v1/W19-4219>
- [28] Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997). <https://doi.org/10.1162/neco.1997.9.8.1735>

- [29] Wells, J.C.: *Accents of English* vol. 3. Cambridge University Press, Cambridge, England, UK (1982). <https://doi.org/10.1017/CBO9780511611759>
- [30] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: *Adv. Neural Inf. Process. Syst.*, vol. 30, 12p. (2017)
- [31] van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**(86), 2579–2605 (2008)