Audiovisual semantic congruency effect with onomatopoeia

Antonio Rei Fidalgo
School of Psychology
University of East London
London, UK

Kohshe Takahashi
School of Psychology
Chukyo University
Nagoya-shi,
Aichi, Japan

Aiko Murata
Faculty of Science and Engineering
Waseda University
Tokyo, Japan

Katsumi Watanabe
Faculty of Science and Engineering
Waseda University
Tokyo, Japan

***Abstract*— It has been reported that when a congruent natural sound precedes briefly presented visual stimuli, it promotes performance in psychophysics detection tasks. Onomatopoeias refer to words that phonetically mimic or suggest actual sounds. Onomatopoeic words are a form of sound symbolism and are frequently used in Japanese language. In this study, we examined whether the presentation of spoken Japanese onomatopoeia to Japanese native-speakers results in visual detection sensitivity changes. Results indicate that when onomatopoeias are presented 227 ms before a visual stimulus, they have a modulatory audiovisual effect. This effect is closer to the results observed with natural sounds than spoken words, with *d*' being lower for onomatopoeias when compared with natural sounds. Such suggests that Japanese spoken onomatopoeias may be processed in a manner that is closer to natural sounds than spoken words and points to behavioral consequences of sound symbolism.**

Keywords — Onomatopoeia; Spoken word; Semantic congruency; Audiovisual interaction; Visual detection.

I. INTRODUCTION

When two sensory modalities are presented concurrently, they often interact with each other [1]. This interaction results in several perceptual and behavioral consequences including, but not limited to, crossmodal enhancement [2], crossmodal attentional shift [3], and crossmodal illusions [4, 5].

Previous studies that scrutinized the effects of presenting auditory and visual stimuli alongside their impact on detection and recognition performances found that the audiovisual semantic congruency effect is depended on the type of sound, audiovisual asynchrony and the experimental task [6, 7, 8, 9, 10, 11]. When participants were asked for a simple visual detection of congruent and incongruent visual stimuli, and sounds were played 346 ms before the visual stimuli, performance was enhanced by natural sounds congruent (ie. matched) with visual stimuli (Experiment 4A, [7]). The same effect was not observed with spoken words that were congruent with visual stimuli (Experiment 4A, [7]). This discrepancy provides an interesting case for examining the long-lasting debate around the relationship between sound and spoken language.

One of the main linguistics theories states that there is virtually no link between the acoustic representation of words and the objects they refer to [12]. This notion effectively rejects "sound symbolism", which refers to the idea and/or situation that spoken sounds contain meaning themselves [13, 14]. However, there are boundary cases. Onomatopoeic words or onomatopoeias phonetically mimic or suggest actual sounds and situations [14]. It has been proposed that onomatopoeia might be considered as a form of sound symbolism through imitation [13, 15].

By employing the protocol set out in Experiment 4A [7] we set out to study whether the auditory presentation of onomatopoeia would modulate visual sensitivity and whether those gains would be more similar to either natural sounds or spoken words. If the spoken onomatopoeias lead to an audiovisual congruency effect closer to natural sounds, it would suggest that the sound symbolism triggered by onomatopoeic words does have some behavioral consequences.

II. METHODS

Participants: Thirty-nine Japanese participants (21 males and 18 females; Age Range 18-28; Mean age: 21.79 ± 1.99) took part in the main study. Each participant provided written informed consent. The study was approved by the Ethics Committee of the University of Tokyo and run in accordance with the Declaration of Helsinki. All reported normal or

corrected-to-normal vision and hearing and, were naive with regards to the purpose of the study.

Apparatus and stimuli: The visual stimuli were presented on a 19-in CRT monitor (Sony G420; 75 Hz refresh rate). The participants sat 57-cm away from the monitor in a dimly-lit experimental chamber. Each trial consisted of the presentation of a target picture, a pattern mask, and a sound stimulus (*Figure* 1). Visual stimuli were presented in black (0.2 cd/m^2) at the center of a white (72.02 cd/m^2) background. The visual targets consisted of 30 outline-drawing figures taken from [16]. The pattern mask consisted of random curves and encompassed an area large enough to completely cover the target picture. The sounds were monaurally sampled at 22050 Hz (8 bit) and presented through closed-ear headphones (HDA 200, Sennheiser). The sound level was far above minimum hearing threshold and set at a comfortable level for each participant.

We used three types of sound stimuli: natural sounds, spoken words, and spoken onomatopoeias. All natural sound files were downloaded from www.findsounds.com (on 12/04/2012). Spoken words and onomatopoeia sounds were produced by recording a female Japanese native speaker (*Table* 2). The sounds were formatted on Audacity (www.audacityteam.org/) to play from the beginning of the sound and to last for 360 ms with equalized peak amplitudes. The sounds were either semantically congruent or incongruent with a presented picture. Congruent auditory stimuli matched the pictures, for example, the picture of a cat was matched to the sound of or the onomatopoeia representing a cat meowing. The incongruent auditory stimuli were randomly chosen from those belonging to a different semantic category from the picture.

Procedure: Like in experiment 4A reported by Chen and Spence [7], participants were shown the visual stimuli before undertaking the experimental task. All 30 pictures were presented for 1000 ms with 500 ms inter-stimulus intervals. The order of picture presentation was randomized across participants. Subsequently, participants were also familiarized with the experimental task with 8-practice trial at 27 ms visual target and subsequently a 16-practice trial at 13 ms visual target with sound stimuli of all three sound conditions taken in a random manner [7]. No data was recorded during familiarization.

In the main session, each trial started by pressing the space key. A blank white screen was presented for 494 ms, which followed by a picture stimulus (picture-present trial) or a blank stimulus (picture-absent trial) for 13 ms. The pattern mask appeared until the participant responded or for 4987 ms immediately afterward in all trials. The sound onset was set 227 ms before the onset of the visual target (*Figure* 1). The participants were instructed to enter 'F' if they saw a picture or 'J' if they did not. The sound-visual onset asynchrony chosen for its consistent higher d' (above 2) and was determined by separate exploratory tests to replicate the result pattern of Experiment 4A in Chen and Spence [7].

The three sound conditions (natural sound, spoken word, and onomatopoeia) were counterbalanced and organized in blocks, thus the main session consisted of three testing blocks. Each block consisted of 120 trials. In each experimental session, all 30 pictures were presented twice, once with congruent sound and once with incongruent sound. The other 60 trials were picture-absent trials. The stimuli were presented in a randomized order within a block, and the order of the three sound conditions were counterbalanced across participants [7].

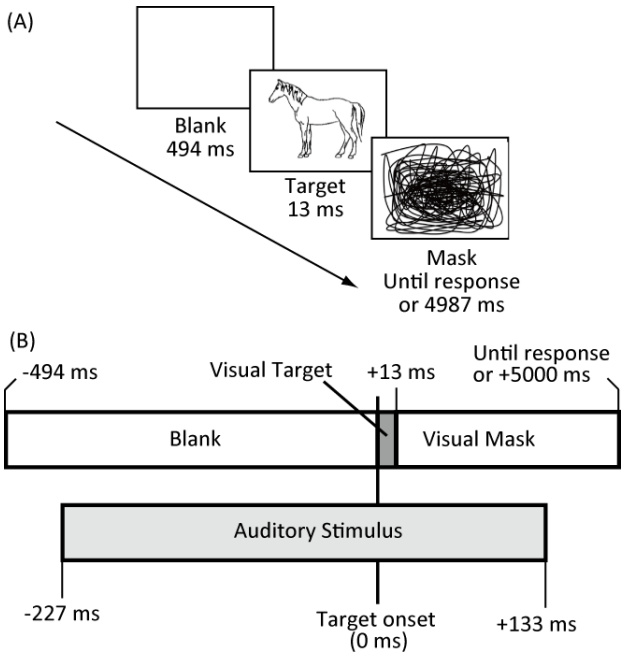

Figure 1. (*A*) The sequence of visual stimuli in each trial: a blank, a target picture (e.g. horse), and a pattern mask. (*B*) A schematic representation of temporal relation between the target picture (visual stimulus) and the sound stimulus.

Data analysis: We calculated sensitivity (*d*') and decision criteria (β) based on signal detection theory [17, 18] for each sound condition and for each sound congruency. The picture-present trials were regarded as signal, while the picture-absent trials were regarded as noise. Sensitivity (*d*') refers to the difficulty in discriminating signal (i.e. picture-present) from noise (i.e. picture-absent) with a lower *d*' score indicating a higher difficulty. Decision criteria (β) address the likelihood of participants selecting 'J' (i.e. picture-absent) as their default answer (β larger than 1 indicates that the participants are biased to choose a "J" response). For statistical analyses, planned t-tests with necessary corrections for multiple comparisons and conventional ANOVAs were independently performed. Conclusions are mainly based on t-tests whereas ANOVA results were taken as indicative rather than determining and conducted mainly for future

comparisons with the previous study and future studies (e.g. for meta-analyses).

III. Results

	Condition	Hit Rate	FA rate	*d*'	*β*	% Correct
Natural Sounds	Congruent	0.790	0.082	2.780	4.421	0.875
	Incongruent	0.756	0.082	2.634	4.386	0.864
Onomatopoeia	Congruent	0.767	0.098	2.467	2.856	0.857
	Incongruent	0.722	0.098	2.281	3.051	0.842
Spoken Words	Congruent	0.739	0.077	2.549	4.767	0.861
	Incongruent	0.719	0.077	2.531	4.321	0.854

Table 1. Aggregated results for the different sound conditions. Our results suggest the presentation of congruent natural sounds that match a corresponding picture enhances the participant's ability to detect briefly presented visual. Such was not observed in the congruent Spoken word condition. Crucially, our experiment also revealed that the onomatopoeia condition has modulatory effects.

One participant was excluded from data analysis due to having a hit rate of less than 0.05. With the remaining participants, a two-way repeated measures ANOVA on *d*' (*Table* 1) showed a significant main effect of congruency [$F(1,37) = 6.71$, MSE = 0.77, $p = 0.014$, partial $\eta^2 = 0.15$] but not for sound condition [$F(2,74) = 2.57$, MSE = 2.11, $p = 0.083$, partial $\eta^2 = 0.06$] with no significant interaction [$F(2,74) = 2.05$, MSE = 0.15, $p = 0.14$, partial $\eta^2 = 0.05$]. Planned pairwise comparison with paired t-tests with Bonferroni correction [$\alpha = 0.05/3 = 0.0167$] revealed that d' in the congruent condition was significantly higher than in the incongruent condition for natural sound [$t(37) = 2.21$, $p = 0.0166$, Cohen's d = 0.73; with Bonferroni correction, one-tailed test was used because the existing hypothesis was that congruency would lead to better performance] and onomatopoeia conditions [$t(37) = 2.90$, $p = 0.003$, Cohen's d = 0.95] while they were comparable in the spoken word condition [$t(37) = 0.24$, $p = 0.404$, Cohen's d = 0.08]. In the natural sound condition, the participants tended to detect the visual stimuli better than in the onomatopoeia condition [$t(37) = 2.19$, $p = 0.035$, Cohen's d = 0.72 with Bonferroni correction; here a two-tailed test was used because there was no hypothesis]. Other comparisons did not reach a statistically significant difference namely, onomatopoeia vs. spoken word [$t(37) = 1.11$, $p = 0.27$, Cohen's d = 0.36] and; natural sound vs. spoken word [$t(37) = 1.20$, $p = 0.24$, Cohen's d = 0.40] (*Figure* 2). Finally, syllable duplication did not impact on participants d' when compared to onomatopoeia without syllable duplication $t(1.020) = 25$, $p = 0.318$.

As for *β* (*Table* 1), no significant statistical differences were found for sound condition [$F(2,74) = 2.00$, MSE = 58.9, $p = 0.15$, partial $\eta^2 = 0.05$], congruency [$F(1,37) = 0.36$, MSE = 0.50, $p = 0.55$, partial $\eta^2 = 0.01$], or their interaction [$F(2,74) = 1.51$, MSE = 2.00, $p = 0.23$, partial $\eta^2 = 0.04$] (*Figure* 2).

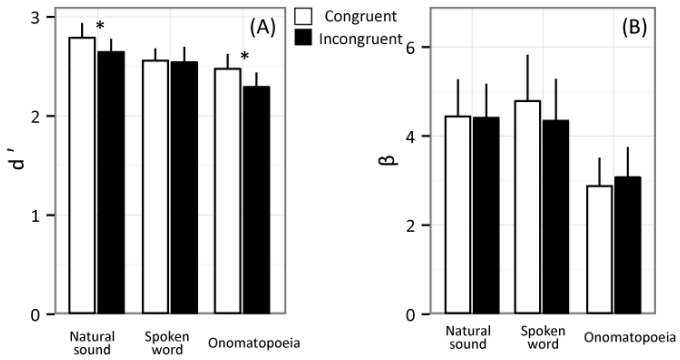

Figure 2. The results of (*A*) sensitivity (d') and (*B*) response bias (β). The error bars indicate a Loftus-Masson 95% confidence interval [19]. Asterisks indicate where significant differences were found in t-tests with Bonferroni correction ($\alpha = 0.05/3 = 0.0167$).

IV. Discussion

In this study, we replicated previous findings by Chen and Spence [7] that, given a particular auditory-visual onset asynchrony, presenting congruent natural sounds that match a corresponding picture enhances the participants' ability to detect briefly presented visual stimuli that are backward masked (with *d*' being consistently higher in congruent trials). Such was not observed in the congruent spoken word condition. Crucially, our experiment also suggests that the onomatopoeia condition has modulatory effects. The effect of the natural words and onomatopoeias were not entirely identical as overall *d*'s were higher in the natural sound condition than in the onomatopoeia condition. These findings might indicate that onomatopoeias are easier to process and/or less distractive when detecting the visual stimuli, but the precise cause of this advantage has to be investigated in the future. Moreover, statistical analyses suggest that both natural sounds and onomatopoeias were effective in inducing

the semantic congruency effect. Thus, our results suggest that detection performance of Japanese onomatopoeia is closer to natural sounds than Japanese spoken words, at least, with the audio-visual temporal asynchrony used in this experiment.

Chen and Spence [7] proposed the audiovisual semantic network (similar to the previously proposed model by [20] where natural sounds and spoken words access the same semantic representation but through different pathways. Specifically, natural sounds may access semantic representations directly (as pictures) but spoken words may access them only after accessing the lexical representation first (as written words) [20, 21, 22]. This model explains why natural sounds enhance the detection of visual stimuli while spoken words do not. This could be observed in our experiment when sound preceded images by 227 ms, as well as in Experiment 4A of Chen and Spence [7] when sound preceded images by 346 ms. Despite the observed effects between the planned comparisons, we also realize the lack of interaction in the statistical analyses and the relatively small effect sizes in the present study. This might be due to the additional (onomatopoeia) condition. While the present results suggest the particularity of onomatopoetic sounds, further temporal optimizations, possibly for each item, might help obtain a clearer pattern in a future study.

Different languages have different onomatopoeic words that cannot be heard as representing the objects they refer to (for example, the sound of a pig would be '*bu*-' /ˈbɯː/ for Japanese; '*oink*' /ˈɔɪŋk/ for English). Onomatopoeic words in one language may share some phonetic or acoustic characteristics with those from other languages [13, 15] and therefore be partly effective across languages. However, the transferability of the semantic congruency effect with onomatopoeia might be observed only across a limited number of languages. For example, the sound of using a pair of scissors is pronounced (and heard) as *su-su* in Chinese, *cri-cri* in Italian, *riqui-riqui* in Spanish, *terre-terre* in Portuguese, *katr-katr* in Hindi, *choki-choki* in Japanese. Moreover, some shape-sound associations (*e.g.* Bouba-Kiki effect) have been implied as being phylogenetic rather than cultural [23]. In relation to this study, it must be noted that the frequency of onomatopoeia use in everyday life varies among languages, whereby Japanese has an exceptionally large number of onomatopoeia that are commonly employed in everyday spoken language and written context [24]. These questions warrant further cross-cultural comparisons.

A previous study has also implied that onomatopoeic sounds can serve as a bridge between spoken words and natural sounds [25]. Neuroimaging data indicate that onomatopoeia and spoken words lead to greater activity in the superior temporal sulcus when compared to natural sounds [25]. In addition, onomatopoeias and natural sounds showed greater activity than spoken words in the left inferior frontal gyrus, an area that is associated with nonverbal processes [25]. Thus, it appears that onomatopoeic sounds are processed by extensive brain regions responsible for processing both verbal and nonverbal sounds [25]. Perhaps, it is the nonverbal process that yielded the semantic congruency effects in visual detection. Future studies on the neural basis of onomatopoeia and their relation to semantic congruency effect would be of interest.

On the other hand, the audiovisual semantic congruency effect might occur through more implicit processes (which would require further experiments). Such implicit processes also might explain why, when hearing onomatopoeia, participants were more likely to miss the visual target (lowering d') and falsely report the presence of the target (decreasing β). We might assume that this change happens because onomatopoeias would trigger greater brain activation compared with spoken words and natural sounds [25, 26], and therefore leave a limited amount of cognitive resources for visual processing of the target (this would warrant further empirical tests that fall outside the scope of this article). In other words, the processing resources might be directed to the onomatopoeia sounds rather than the visual stimulus, which might result in the lower d' and β.

Given our efforts to replicate as close as possible experiment 4A from Chen and Spence [7] our study did not contain a no-sound or white noise condition. Such condition would constitute an informative control in order to determine whether the congruency effect stems from a facilitating effect of congruent sounds or rather from an interference effect due to the incongruent sound. Moreover, preliminary experiments conducted for this study where all three sound conditions were presented in the same block (i.e. there would be an equal likelihood of natural sounds, spoken words or onomatopoeias being presented), resulted in no statistically significant differences between the different stimuli and therefore indicate that this modulatory effect might not be totally automatic or bottom-up.

In conclusion, we report that when sound preceded visual stimuli by 227 ms, the presentation of congruent natural sounds and Japanese onomatopoeias modulates the detection of briefly presented visual stimuli. The same was not observed with congruent spoken words. These results suggest that Japanese onomatopoeic words are processed more similarly to natural sounds than spoken words. Such happens despite onomatopoeias being clearly identified as spoken words rather than natural sounds. It warrants further investigation whether this holds true for other languages. In a similar manner, it would also be interesting to conduct a neuroimaging study to investigate how onomatopoeias are represented and how they interact with other language functions in the brain.

ACKNOWLEDGMENT

This study was partly supported by a Japan Society for the Promotion of Science post-doctoral fellowship to ARF and;

CREST, Japan Science and Technology Agency and KAKENHI grants (24300279 to KW and 25700013 to KT). The sponsors played no role in study design, data collection and analyses and interpretation of the data, writing and submission of the article.

REFERENCES

[1] G. Calvert, C. Spence, and B.E. Stein (Eds.), "The handbook of multisensory processes," Cambridge, Mass: MIT Press, 2004.

[2] B.E. Stein, and T.R. Stanford, "Multisensory integration: current issues from the perspective of the single neuron," Nat. Rev. Neurosci., vol. 9, pp. 255–266, 2008.

[3] C. Spence, and J. Driver (Eds.), "Crossmodal space and crossmodal attention," Oxford: Oxford University Press, 2004.

[4] I.P. Howard, and W.B. Templeton , "Human spatial orientation," New York: Wiley, 1996.

[5] H. McGurk, and J. MacDonald, "Hearing lips and seeing voices," Nature vol. 264, pp. 746–748, 1976.

[6] Y.-C. Chen, and C. Spence, "When hearing the bark helps to identify the dog: semantically-congruent sounds modulate the identification of masked pictures," Cognition, vol. 114, pp. 389–404, 2010.

[7] Y.-C.Chen, and C. Spence, "Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity," J Exp Psychol Hum Percept Perform, vol. 37, pp. 1554-1568, 2011.

[8] Y.-C. Chen, and C. Spence, "The time-course of the cross-modal semantic modulation of visual picture processing by naturalistic sounds and spoken words," Multisens Res, vol. 26, pp. 371–386, 2013.

[9] Y.-C. Chen, P.-C. Huang, S.-L. Yeh, C. and Spence, "Synchronous sounds enhance visual sensitivity without reducing target uncertainty," Seeing Perceiving, vol. 24, pp. 623–638, 2011.

[10] Y.-C. Chen, S.-L. Yeh, and C. Spence, "Crossmodal constraints on human perceptual awareness: auditory semantic modulation of binocular rivalry," Front Psychol, vol. 2, pp. 212, 2011.

[11] J.-Y. Hsiao, Y.-C. Chen, C. Spence, and S.-L.Yeh, "Assessing the effects of audiovisual semantic congruency on the perception of a bistable figure," Conscious Cogn, vol. 21, pp. 775–787, 2012.

[12] F. De Saussure, "Course in General Linguistics," New York: Columbia University Press, 2011.

[13] R. Brown, "Psycholinguistics," in: Selected Papers by Roger Brown, Free Press, pp. 258–273, 1970

[14] P.H. Matthews, "The concise Oxford dictionary of linguistics," Third edition. ed, Oxford: Oxford University Press, 2014.

[15] M.F. Assaneo, J.I. Nichols, and M.A. Trevisan, "The anatomy of onomatopoeia," PLoS ONE, vol. 6, pp. e28317, 2011.

[16] J.G. Snodgrass, and M. Vanderwart, "A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity," J Exp Psychol Hum Learn, vol. 6, pp. 174–215, 1980.

[17] D.M. Green, and J.A. Swets, "Signal Detection Theory and Psychophysics," New York: Wiley, 1966.

[18] N.A. Macmillan, "Detection theory: a user's guide," 2nd ed. ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2005.

[19] T. Baguley, "Calculating and graphing within-subject confidence intervals for ANOVA," Behav Res Methods, vol 44, pp. 158–175, 2012.

[20] W.R. Glaser, and M.O. Glaser, "Context effects in stroop-like word and picture processing," J Exp Psychol Gen, vol. 118, pp. 13–42, 1989.

[21] A. Cummings, R. Ceponiene, A. Koyama, A.P. Saygin, J. Townsend and F. Dick, "Auditory semantic networks for words and natural sounds," Brain Res., vol. 1115, pp. 92–107, 2006.

[22] M. Coltheart, "Dual routes from print to speech and dual routes from print to meaning. Some theoretical issues", In A. Kennedy, (Ed.), Reading as a Perceptual Process, North Holland, 2000, pp. 475–490.

[23] A.J. Bremner, S. Caparos, J. Davidoff, J. de Fockert, K.J. Linnell, and C. Spence, " "Bouba" and "Kiki" in Namibia? A remote culture make similar shape-sound matches, but different shape-taste matches to Westerners," Cognition, vol. 126, pp.165–172, 2013.

[24] S. Hamano, "The sound-symbolic system of Japanese, Studies in Japanese linguistics," Kurosio. 1998.

[25] T. Hashimoto, N. Usui, M. Taira, I. Nose, T. Haji, and S. Kojima, "The neural mechanism associated with the processing of onomatopoeic sounds," Neuroimage, vol. 31, pp. 1762–1770, 2006.

[26] N. Osaka, "Walk-related mimic word activates the extrastriate visual cortex in the human brain: an fMRI study," Behav. Brain Res., vol. 198, pp. 186–189, 2009.

APPENDIX

	Japanese Word	Japanese Onomatopoeia
Bird	Tori	chunchun
Cat	Neko	nyaanyaa
Cow	Ushi	mouu
Dog	Inu	wanwan
Duck	Ahiru	guwaguwa
Eagle	Washi	kyuu
Elephant	Zo	paoo
Fly	Ae	buun
Frog	Kairu	kerokero
Goat	Yagi	mee
Horse	Uma	hiin
Mouse	Nezumi	chii
Pig	Buta	buubuu
Rooster	Niwatori	kokekokkoo
Tiger	Tora	gaoo
Cannon	Taiho	bon
Car	Kuruma	buun
Door	Doa	batan
Drum	Taiko	dondon
Guitar	Guita	juun
Gun	Kenju	bakyuun
Hammer	Kanazushi	tonton
Lock	Kagi	kacha
Motorcycle	Otobai	guun
Piano	Piano	tintin
Scissors	Kazami	jokijoki
Switch	Suitchi	pachi
Telephone	Denwa	pururuun
Trumpet	Trumpeto	pararira
Whistle	Fue	pii

Table 2. Stimuli Pronunciation: the English words are followed by the Japanese pronunciation (using western characters) and the Japanese onomatopoeia