Supplemental Material

No Evidence That Sound-Shape Associations Influence Temporal Resolution in Humans: Five Non-Replications of Parise and Spence (2009) and Meta-Analyses

Suddha Sourav, Brigitte Röder, Franka Ambsdorf, Andromachi Melissari, Miketa Arvaniti, & Argiro Vatakis

S1. Frequency Domain Characteristics of the Speech Stimuli in Experiment 4

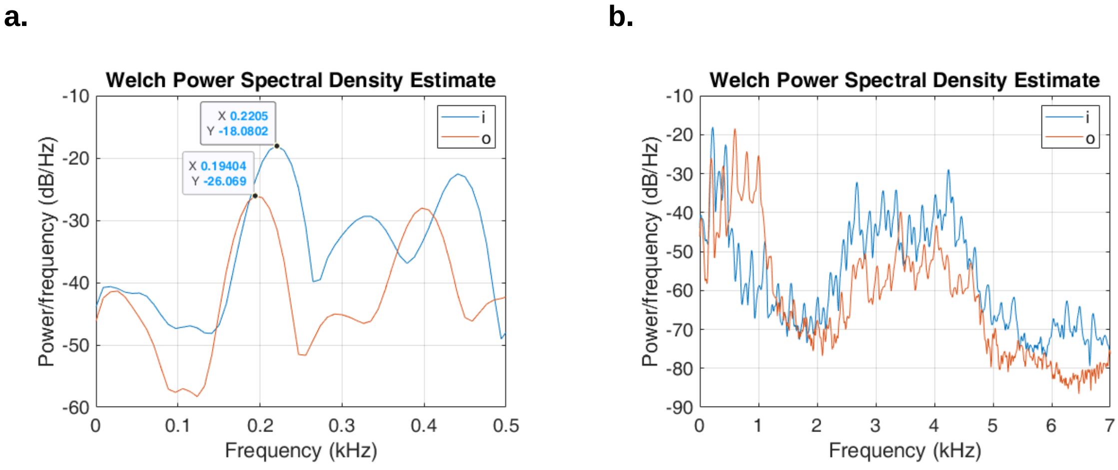
In experiment 4, the auditory stimuli were the phonemes /o/ and /i/ which were articulated by a female voice for a duration of 200 ms. To analyze the frequency domain characteristics of the phonemes, we derived the frequency spectra (i.e., the power spectral density estimates) of the audio signals using the Welch method (sampling rate = 44.1 kHz, 1700 points, 50% overlap, see Fig. S1).

The range of fundamental frequency (F_0) of speech has been reported to be about 165 – 255 Hz for human female adults (Fitch & Holbrook, 1970). In experiment 4 of the present study, the F_0 of the vowel /o/ was 194 Hz, and the F_0 for the vowel /i/ was 221 Hz (Fig. S1a). Moreover, the frequency spectra of the vowel sounds indicated that the vowel sound /i/ had a higher power spectral density in the high-frequency speech range (2 – 7 kHz) compared to /o/ (Fig. S1b).

In addition to F_0, peaks in the frequency spectral envelope of vowel sounds, known as *formants*, have been argued to play a critical role in sound-shape associations (Ćwiek et al., 2021). In particular, the association between the phoneme /i/ and spiky visual shapes has been hypothesized to be driven by its considerably high second formant frequency (F_2). Here, we additionally calculated the formant frequencies of the phonemes in experiment 4 using the *formant estimation toolbox* in MATLAB R2022a (Rabiner et al., 2023; The MathWorks Inc., 2022). The parameters for formant estimation were as follows: linear predictive code system order = 32, frame length = 40 ms, frame shift = 10 ms, minimum valid pole radius = 0.9, and formant distance threshold = 0.05. The analysis indicated that the mean estimates for the first and the second formants (F1, F2) for /o/ were (394 Hz, 892 Hz), and for /i/, (299 Hz, 2841 Hz). These values agree well with canonical formant frequency ranges reported for these vowel sounds (Peterson & Barney, 1952), and provide a quantitative corroboration of the observation that the phoneme /i/ had a higher power spectral density at higher speech frequencies compared to the phoneme /o/.

Figure S1

Frequency Spectra of the Vowels Sounds in Experiment 4



Note. Power spectral density estimate for the vowel sounds /i / and /o/ used in experiment 4, calculated with the Welch method (sampling rate = 44.1 kHz, 1700 points, 50% overlap). **a.** The fundamental frequency for the phoneme /i/ was approximately 221 Hz, which was higher than that of /o/ (approximately 194 Hz). Note that the x – axis has the unit of kHz. **B.** The same plot with the x-axis ranging from 0 – 7 kHz. Higher power spectral density in the high-frequency speech range (2 – 7 kHz) can be observed for the vowel /i/ compared to the vowel /o/.

S2. Sound-*Size* Association and Temporal Resolution: A Conceptual (Non-) Replication of Experiment 1 of Parise and Spence (2009)

In the preregistered experiment 3 of the present study, we aimed to investigate whether sound-shape association (SSA) modulates multisensory temporal resolution as reported in experiment 2 of Parise and Spence (2009), and whether such a modulation might be observable selectively for smaller visual stimuli. For this analysis, the factor *congruence* was defined based on SSA, that is, the constituent visual stimulus shapes and sound pitch. For example, the combination of a spiky visual shape with a higher-pitched tone was coded as a congruent SSA combination, whereas a stimulus with a spiky visual shape and a lower-pitched tone was coded as an incongruent one (Table S2). The additional factor *size* allowed us to investigate whether visual stimulus size might modulate a hypothetical SSA congruence effect. We found no evidence that SSA modulated temporal resolution, irrespective of visual stimulus size (see *Results: No Replication of Parise and Spence (2009) with Identical Nonparametric Tests* in the article).

Parise and Spence (2009) additionally reported that an association between auditory pitch and visual stimulus *size* (sound-size association) modulated multisensory temporal resolution akin to sound-shape association in the first experiment of the article. Replicating this experiment was not a central aim of our preregistered experiment 3. However, as we used different visual stimulus sizes, the data of this experiment allowed a conceptual replication analysis of experiment 1 of Parise and Spence (2009). For this analysis, the factor *congruence* was coded not based on the visual stimulus *shape* and the auditory stimulus pitch, as for the SSA analysis, but based on the visual stimulus *size* and the auditory stimulus pitch (Table S2).

Table S2

Visual and Auditory Stimulus Properties in the Preregistered Experiment 3 and the Operationalization of Sound-Shape/Sound-Size Association

Visual Stimulus Size	Visual Stimulus Shape	Auditory Stimulus Pitch	Sound-Size Association Congruence	Sound-Shape Association Congruence
Small	Smooth	Low	0	1
Small	Smooth	High	1	0
Small	Spiky	Low	0	0
Small	Spiky	High	1	1
Large	Smooth	Low	1	1
Large	Smooth	High	0	0
Large	Spiky	Low	1	0
Large	Spiky	High	0	1

Note. For the sound-shape/sound-size congruence, 1 denotes a congruent, and 0 an incongruent combination.

Collapsing across the visual stimulus shapes, we found no evidence that sound-size congruence modulated multisensory TOJ performance, Wilcoxon signed-rank test, median

Δ_{JND} (IQR) = 1.16 (−9.40, 18.61) ms, Z = 0.511, p = .610. This result once again highlights the need for evidence synthesis and indicates that multisensory temporal resolution may be unaffected by higher-level associations beyond SSA.

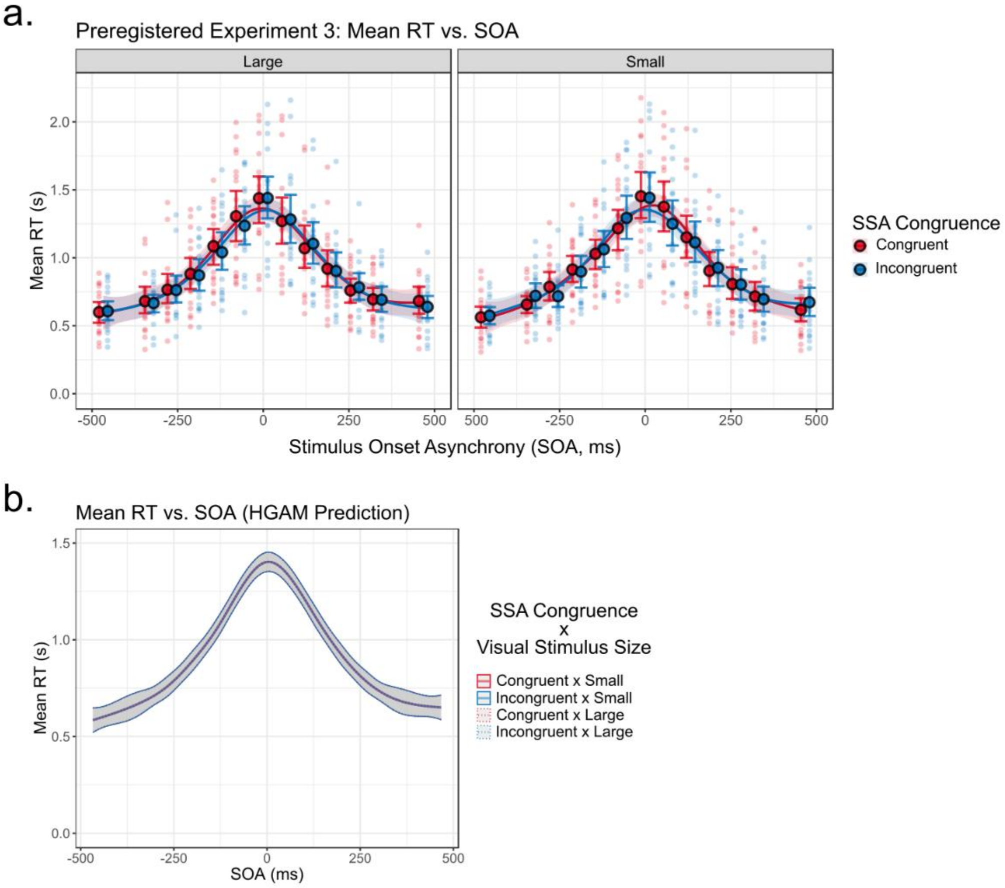
It is to be noted that Parise and Spence (2009) used circular visual stimuli respectively subtending visual angles of 2.1° and 5.2° as the smaller and the larger visual stimuli, whereas in our study, the two visual stimulus shapes employed for investigating SSA effects, each subtending 5.2°/10.4°, were used. If sound-size associations meaningfully modulated temporal resolution, we would expect to observe a generalization across different visual stimulus shapes, as well as a replication of this effect for a range of visual stimulus sizes.

S3. No Evidence That SSA Modulates Reaction Times in a Temporal Order Judgement Task

We tested whether SSA modulated reaction times (RT) for the temporal order judgement (TOJ) task in the preregistered experiment 3 (https://aspredicted.org/ip5ht.pdf). RTs were recorded with response buttons connected to a low-latency and low-jitter data acquisition system (*NIDAQ – 6509*, National Instruments, Austin, TX, USA). As in the main analysis, RTs less than 100 ms and greater than 3.0 s were discarded from the analysis. The RT vs. SOA plots revealed a characteristic slowing down of the RTs for lower SOAs (Fig. S3a). With a hierarchical generalized additive model (HGAM, Pedersen et al., 2019), we tested whether the RT vs. SOA curves were different across the different SSA conditions (congruent/incongruent) and visual stimulus sizes employed in the study (5.2°/10.4°). The general shape of the RT vs. SOA curves was modeled by utilizing a thin plate smooth spline with a loss function that aimed to minimize the absolute sum of squared values of the residuals and contained a penalty for the complexity (*"wiggliness"*) of the curve as well (Wood, 2017). In addition to the general RT vs. SOA curve, additional penalized thin plate smooth splines modeled the group-level differences to this general shape, one for each *Congruence* × *Visual Stimulus Size* condition. These group-level smooth splines were constrained to share the same wiggliness parameter (cf. Pedersen et al., 2019, "*model GS*"). The HGAM was fitted with the *R* library *mgcv* v1.8-42 running on *R* v4.2.2 (R Core Team, 2016; Wood, 2011). The result of the hierarchical model indicated that SOA had a nonlinear effect on RT, smooth term for SOA, $F(8.172, 926.828) = 90.65$, $p < 2\times10^{-16}$ (Fig. S3b). In contrast, we found no evidence that any of the specific *Congruence* × *Visual Stimulus Size* conditions differed from this general shape, smooth term for *Congruence* × *Visual Stimulus Size*, $F(1.084\times10^{-5}, 926.828) < 0.01$, $p = .989$. In Fig. S3b, the predictions of the HGAM model for the four conditions are plotted.

Figure S3

Dependence of Reaction Times on Stimulus Onset Asynchrony and Sound-Shape Association in the Preregistered Temporal Order Judgement Task



Note. **a.** Plot of mean reaction times (RT) against stimulus onset asynchronies (SOA) for congruent and incongruent SSAs in the preregistered experiment 3, $n = 18$, left panel: audiovisual presentations with a larger (10.4°) visual stimulus, right panel: audiovisual presentations with a smaller (5.2°) visual stimulus. Curves are separate generalized additive model estimates for each congruence and visual stimulus size, and the ribbons indicate 95% confidence intervals of the mean estimates. At each SOA, filled large circles indicate the mean RT, and the error bars show 95% confidence intervals of the means obtained by smoothed bootstrapping with Gaussian kernels (Wołodźko, 2019). Individual data points have been jittered for readability. **b.** Hierarchical generalized additive model (HGAM) estimates for the four SSA *Congruence* × *Visual Stimulus Size* conditions. Grey ribbons indicate 95% confidence intervals of the mean. A single curve is perceived due to the four estimates virtually completely overlapping.

Previously, in an implicit association task, Parise and Spence (2012) reported slower RTs to stimuli containing incongruent compared to congruent SSA, a finding that has been repeatedly replicated after its discovery (Peiffer-Smadja & Cohen, 2019; Silva & Bellini-Leite, 2020). The current results argue that unlike implicit association tasks, TOJ task performance is not modulated by SSA. These behavioral results are consistent with the hypothesis that SSA does not lead to obligatory and general modulations of behavioral performance, but rather might do so in a task-specific manner.

References

Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., Kawahara, S., Koutalidis, S., Krifka, M., Lippus, P., Lupyan, G., Oh, G. E., Paul, J., Petrone, C., Ridouane, R., Reiter, S., Schümchen, N., Szalontai, Á., Ünal-Logacev, Ö., … Winter, B. (2021). The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1841), 20200390. https://doi.org/10.1098/rstb.2020.0390

Fitch, J. L., & Holbrook, A. (1970). Modal Vocal Fundamental Frequency of Young Adults. *Archives of Otolaryngology*, *92*(4), 379–382. https://doi.org/10.1001/archotol.1970.04310040067012

Parise, C. V., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: A study using the implicit association test. *Experimental Brain Research*, *220*(3), 319–333. https://doi.org/10.1007/s00221-012-3140-6

Pedersen, E. J., Miller, D. L., Simpson, G. L., & Ross, N. (2019). Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ*, *7*, e6876. https://doi.org/10.7717/peerj.6876

Peiffer-Smadja, N., & Cohen, L. (2019). The cerebral bases of the bouba-kiki effect. *NeuroImage*, *186*, 679–689. https://doi.org/10.1016/j.neuroimage.2018.11.033

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184. https://doi.org/10.1121/1.1906875

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Rabiner, L., Schafer, R., Vedula, K., & Yedithi, S. (2023). *Formant Estimation* [MATLAB]. MATLAB Central File Exchange. https://de.mathworks.com/matlabcentral/fileexchange/45315-formant-estimation

Silva, D. M. R., & Bellini-Leite, S. C. (2020). Cross-modal correspondences in sine wave: Speech versus non-speech modes. *Attention, Perception, & Psychophysics*, *82*(3), 944–953. https://doi.org/10.3758/s13414-019-01835-z

The MathWorks Inc. (2022). *MATLAB* (9.12.0.1927505 [R2022a] Update 1) [Computer software]. The MathWorks Inc.

Wołodźko, T. (2019). *Kernelboot: Smoothed bootstrap and random generation from kernel densities.* (0.1.6) [Computer software]. https://cran.r-project.org/package=kernelboot

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279