Pitch Expectancy Modulates Cross-Modal Correspondence Effect

Lazarina Georgieva (lazarina.georgieva@gmail.com)
Department of Cognitive Science and Psychology, New Bulgarian University
21 Montevideo Blvd., Sofia 1618, Bulgaria

Armina Janyan (ajanyan@cogs.nbu.bg)
Department of Cognitive Science and Psychology, New Bulgarian University
Research Center for Cognitive Science, New Bulgarian University
21 Montevideo Blvd., Sofia 1618, Bulgaria

Abstract

A number of studies have investigated whether cross-modal correspondence effect occurs in purely automatic manner or whether top-down processes can be involved in the processing. The current study addresses the disparity in the research conducting two experiments, using a classical audiovisual cross-modal correspondence paradigm and testing possible involvement of the endogenous component in the effect. Experiment 1 replicated previous findings and showed presence of cross-modal correspondence between pitch and spatial position. However, the effect was significant only in upper spatial position. Experiment 2 showed that task-related pitch probability manipulation made the cross-modal correspondence effect to disappear, however revealing an asymmetrical pattern that was highly dependent on pitch probability and spatial position. Overall, the results suggested a non-automaticity of the cross-modal correspondence effect and a possible involvement of endogenous component in the effect.

Keywords: audiovisual cross-modal correspondence; automaticity; endogenous process; RT

Introduction

People experience life through multiple senses. Whether listening to music while running through the park or looking for a friend in a crowded room, the brain is constantly managing information from multiple sensory modalities. Since the beginning of the 20th century, psychologists and neuroscientists have conducted studies to investigate how external stimuli from different modalities are processed in the brain, with a rise in the interest on the topic in the last couple of decades.

Some studies have explored multisensory integration i.e., the process of selecting, organizing and combining information from different sensory modalities (visual, auditory, touch, smell or taste) into a comprehensive representation (Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010), and others have focused on cross-modal correspondences: associations the brain may construct between the features of stimulus in one modality (e.g., visual) and the features of a stimulus in another modality (e.g., auditory) (Spence, 2011). In addition to investigating the presence of such correspondences, a number of studies have also explored whether cross-modal correspondences occur in an automatic way or they operate at a more strategic level.

Cross-modal correspondences have been defined as the associations between the dimensions or features of stimuli presented in different modalities (Spence, 2011). Cross-modal correspondences can be found between different modalities, but for the purpose of this paper we will solely focus on audiovisual cross-modal correspondences, in particular, between pitch and spatial position of a stimulus.

Audiovisual cross-modal correspondences have been discovered between the pitch of a tone and spatial position: high pitch was associated rather with an object positioned in upper visual field and low pitch – with an object positioned in lower visual field (Evans & Treisman, 2010; Evans, 2020). The association was reflected in reaction time: participants responded to 'congruent' trials faster than 'incongruent' (e.g., high pitch coupled with an object in lower visual field). The effect of cross-modal association seems to be stable and reliable, however, there is little consensus concerning the processes behind audiovisual cross-modal correspondence associations. One of the most debated questions in recent years is whether the associations occur in purely automatic (exogenous, stimulus-driven) manner, or whether top-down, endogenous processes also contribute to them. A seminal paper by Evans & Treisman (2010) as well as later papers (e.g., Parise & Spence, 2012) suggested that the cross-modal correspondence is automatic in nature, takes place at the perceptual level, and is independent on selective attention (Evans, 2020). However, other studies (e.g., Getz & Kubovy, 2018; Chiou & Rich, 2012) using a cuing paradigm, suggest that the cross-modal effect may involve an endogenous component. Authors of a recent study (Janyan et al., 2022) noted that the cuing paradigm gives enough time for an endogenous process to develop thus, they applied simultaneous brief stimuli presentation as in Evans & Treisman (2010). Results of a complicated experiment with different tasks (Janyan et al., 2022) seemingly supported involvement of selective attention in cross-modal correspondence effect.

While Evans (2020) and Janyan et al. (2022) based their theoretical arguments mainly on theories of automaticity (Moors & De Houwer, 2006) and varied either type of perceptual or cognitive load (Evans, 2020) or tasks that would focus on a particular auditory feature ignoring another one (Janyan et al., 2022), we took a slightly different approach. We introduced an undoubtedly endogenous element into the cross-modal correspondence paradigm – that

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

of expectation of a particular task-related stimulus. The combination of top-down (expectation) with bottom-up (stimuli perceptual features) processes relies on a framework, proposed by Tang, Wu & Shen (2016), which suggests that endogenous processes within a multisensory object can spread in both exogenous and endogenous manner. The experiments aimed to test whether the classical cross-modal correspondence effect would be (dynamically) modulated by the integration of the two types of processes or the processes will not interact. Supposedly, the latter case would suggest that the effect is automatic in its classical sense and would agree with earlier studies (Evans & Treisman, 2010; Evans, 2020). Otherwise, if the results show a modulation of the cross-modal effect, it would lead to a suggestion that top-down/endogenous processes are/can be involved in the cross-modal correspondence effect which would, it its turn, support the account put forward by Tang et al. (2016).

Two experiments were conducted that applied the same methodology. Experiment 1 was a replication of a classical cross-modal (audio-visual) correspondence effect, with simultaneous brief stimuli presentation as in Evans & Treisman (2010). In addition, we included a control condition, testing whether the effect is symmetrical across spatial positions or the 'vertical' conditions differentially contribute to the effect as in Janyan et al. (2022). Experiment 2 presented task-related stimulus probability manipulation.[1]

Experiment 1. Replication

Method

Participants Forty-eight students participated in the experiment (7 males, age M(SD)=23.1(7.1) years old). They either volunteered or participated in exchange for a course credit after giving their written informed consent. All participants had normal or corrected to normal vision. None of the participants reported any hearing disability.

Stimuli, Design, and Procedure The design of the experiment was 2 (Pitch: high vs. low) x 3 (Spatial position: center vs. down vs. up) within-subjects design. High (3000 Hz) and low (1000 Hz) pitch stimuli were created using Audacity® v.3.3.2 software as sine tones with length of 100 ms (sample rate: 44100 Hz, bit sample 32). Visual stimulus was a black square presented on a silver background. The stimulus subtending 1.9° visual angle was presented either centrally, or 8.7° above or below the screen center.

Stimuli presentation and response collection was controlled by E-prime 2.0 software (Schneider, Eschman, & Zuccolotto, 2002). Participants were tested individually in sound-proof booth in front of 22-inch monitor with resolution of 1920x1080 pixels and refresh rate of 50 Hz. Participants were positioned at around 57 cm from the screen. Sounds were played binaurally via headphones with approximate intensity of 60 dB. Participants were asked to look at the screen and to categorize a pitch (high/low) as fast and as accurately as possible pressing a corresponding button on the computer keyboard ("J" or "K") with one hand. Response mapping was counterbalanced between participants. A trial included centrally presented fixation cross (500 ms), followed by simultaneous auditory and visual stimuli presentation (100 ms). After stimuli disappearance participants had 1500 ms to respond. Inter-trial interval was 1000 ms.

Participants first went through a familiarization block of 10 trials where they heard sounds presented for 100 ms together with their pitch labels (high or low) that stayed on the screen for 1000 ms. After the familiarization, participants gave start to the practice set of trials by pressing a button on the computer keyboard. The practice block consisted of 12 trials with a feedback after each trial. After the practice, participants started the experimental procedure by pressing a button. The experimental procedure consisted of 156 trials (26 trials per condition). The experimental part was run without the feedback. Trials were pseudorandomized in such a way that there were no more than two consecutive trials of the same condition. The experiment took around 8-10 minutes to complete.

Results and Discussion

Data of twelve participants were excluded. Data of six of them were removed due to low accuracy (< 80%), five of them did not follow instructions to look at the screen and data of one of them were removed due to technical issues.

From the remaining data of 36 participants' erroneous responses (5.48%) and outliers (5.01%) with RT outside of $\pm$ 2SD per participant and per condition means were excluded from the RT analysis. The data were averaged by subject and then entered into repeated measures ANOVA (rANOVA) with pitch and spatial position as within-subject variables. Bonferroni post-hoc test was applied where appropriate. Table 1 presents descriptive statistics per condition.

Table 1: Mean RTs, standard deviations (SD) in parentheses, and 95% confidence intervals (CI) per condition, ms.

Position	High Pitch	Low Pitch
Center	322(84), 294–350	342(93), 310–374
Down	350(90), 319–380	338(101), 330–393
Up	323(80), 296–380	362(94), 331–394

rANOVA obtained a significant main effect of pitch ($F(1,35)=4.92$, $p=0.03$ $\eta_p^2=0.12$), suggesting that participants categorized high pitch (332 ms) faster than low pitch (348

[1] The study was approved prior to the beginning of the experiments by the Ethics Committee of the Department of Cognitive Science and Psychology, New Bulgarian University, Sofia, Bulgaria.

ms). Main effect of spatial position was also significant ($F(2,70)=3.98$, $p=0.02$, $\eta_p^2=0.10$). A post-hoc analysis showed that the center (332 ms) position produced faster responses than the down (344 ms) position ($p=0.036$). There was no difference between control and up position (342 ms, $p=0.074$). Most importantly, a significant two-way interaction ($F(2,70)=9.29$, $p<0.001$ $\eta_p^2=0.20$) was obtained (see Figure 1). A post-hoc comparison showed that participants were significantly faster to respond to high pitch (323 ms), compared to low pitch (362 ms) when the stimulus was positioned in the upper part of the screen ($p=0.002$). suggesting significant cross-modal correspondence effect only for pitch and visual stimuli presented in upper position. Finally, Bonferroni post-hoc test found significant difference between high pitch in lower (350 ms) and upper (323 ms) position ($p=0.03$) as well as between lower and central (322 ms) position ($p=0.02$). Critically, the difference in low pitch central and upper position was not significant ($p>0.1$) but between lower and upper position was ($p=0.028$) (cf. Figure 1).

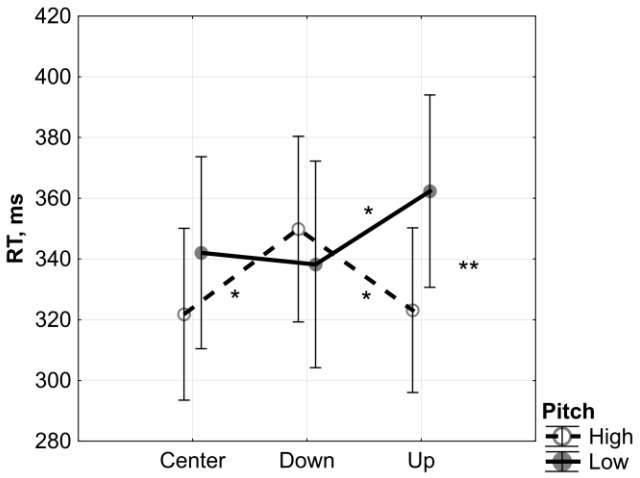

Figure 1: Experiment 1. A spatial position by pitch interaction. Vertical bars denote 95% CI.
* $p<0.05$; ** $p<0.01$.

The results suggested that the cross-modal correspondence effect is accounted for the upper position of the stimulus only. To test whether there was an overall correspondence effect, the data were collapsed into congruent, incongruent, and control conditions (e.g., as in Evans & Treisman, 2010). rANOVA showed a significant main effect ($F(2,70)=11.25$, $p<0.001$, $\eta_p^2=0.24$), with incongruent condition (356 ms) being significantly slower ($p<0.001$) than the congruent (330 ms) and the control (332 ms) one. Thus, the collapsed data showed a congruence effect and masked the lack of contribution of the lower position of the stimulus. A recent study (Janyan et al, 2022) also obtained a cross-modal effect only for the upper position of a visual stimulus. The authors interpreted it relying on two main evidences: First, that there is an asymmetry in saccadic speed between upper and lower visual fields with faster saccades to the upper one (Greene, Brown, & Strauss, 2019), and second, that the natural correlation between high pitch and upper part of visual space is much stronger than the correlation between the low pitch and lower part of the visual space (Parise, Knorre, & Ernst, 2014). Thus, the authors argued that because of these two evidences, the conflict in the upper part of the visual space is much stronger than in the lower part of the visual space and, therefore, observable (Janyan et al., 2022).

Overall, the experiment replicated the cross-modal corresponding effect and suggested that, probably, in some cases it could be useful to test where the effect "comes from" before collapsing the data.

Next experiment directly tested the hypothesis of endogenous processes involvement in cross-modal correspondence effect introducing pitch probability manipulation. The probability manipulation allowed to vary the top-down expectancy about a particular pitch appearance. Two separate experimental blocks were constructed, one with 75% high pitch probability and 25% low pitch probability appearance, and the other one – with 75% low pitch probability and 25% high pitch probability appearance. For convenience, the blocks were called 75% high pitch, and 75% low pitch blocks.

Experiment 2. Pitch Probability Manipulation

Method

Participants Overall, 96 university students participated in the experiment (26 males, age M(SD)=23.9(6.10) years old). Of them, 54 students (age M(SD)=23.6(5.6) years old) participated in high pitch 75% probability block, and 42 students (age M(SD)=24.05(6.7) years old) – in the low pitch 75% probability block. They either volunteered or participated in exchange for a course credit after giving their written informed consent. All participants had normal or corrected to normal vision. None of the participants reported any hearing disability.

Stimuli, Design, and Procedure The design of the experiment was 2 (Pitch: high vs. low) x 3 (Spatial position: center vs. down vs. up) x 2 Pitch probability (75% high vs. 75% low) mixed design. Auditory and visual stimuli were the same as in Experiment 1, as well as the task and the trial procedure. Two separate participants' groups were run addressing the pitch probability manipulation (75% low and 75% high pitch). Each pitch probability block contained 216 experimental trials. Instruction did not mention the pitch probability. After half of the trials (108) participants could have some rest if they wished. Each block took about 12-15 min to complete.

Results and Discussion

Data of 24 participants were removed from the analysis. Data of 13 of those were removed due to low accuracy (<80%, 11 from 75% high pitch probability group), and 11 did not follow instructions to look at the screen (7 from 75% high pitch probability group). Data of 37 participants were

accepted for the analysis in the 75% high pitch probability group, and of 35 -- in the 75% low pitch probability group.

Erroneous responses (4.90%) and outliers (4.85%) with RT outside of ± 2SD per participant and per condition means were excluded from the RT analysis. The data were averaged by subject and then entered into rANOVA with pitch and spatial position as within-subject variables and probability group – as a between-subject variable. Bonferroni post-hoc test was applied where appropriate. Table 2 presents descriptive statistics per condition.

Table 2: Mean (SD), 95% CI per condition, RT, ms. The first column represents pitch probability manipulation.

75%	Position	High Pitch	Low Pitch
	Center	259(89), 229–289	311(80), 285–338
High	Down	275(86), 246–304	309(100), 276–342
	Up	265(90), 235–295	327(104), 293–305
	Center	332(94), 300–364	301(87), 271–331
Low	Down	328(81), 300–356	311(91), 280–342
	Up	313(71), 287–338	339(100), 305–374

rANOVA with pitch (high vs. low), spatial position (center vs. down vs. up) as within-subject factors and pitch probability (75% High vs. 75% Low) as between-subjects factor was conducted. The analysis showed a main effect of pitch ($F(1,70)=20.73$, $p<0.001$. $\eta_p^2= 0.23$) and main effect of spatial position ($F(2,140)=6.26$, $p=0.002$, $\eta_p^2=0.08$). Post-hoc analysis suggested significant difference being between the center and up positions ($p=0.002$), with participants responding faster in trials where the square was presented centrally (301 ms) compared to upper position (311 ms). A main effect of pitch probability group was not significant ($p>0.1$). The analysis also showed a two-way interaction between pitch and spatial position ($F(1,70)=37.64$, $p<0.001$, $\eta_p^2=0.35$). Further, the analysis obtained significant two-way interaction between pitch and position ($F(2,140)=12.651$, $p<0.001$, $\eta_p^2=0.15$). A spatial position by probability group interaction was not significant ($p>0.7$).

Critically, the analysis revealed a three-way interaction between pitch, spatial position and pitch probability group ($F(2,140)=4.74$, $p=0.010$, $\eta_p^2= 0.06$) (see Figure 2). In 75% High pitch group there were significant differences between high and low pitch in all spatial positions: central ($p<0.001$), down ($p=0.015$) and up ($p<0.001$). However, in the 75% Low pitch group significant differences between low and high pitches in the three spatial positions were not found (central ($p=0.07$), down ($p=1.0$), and up ($p>0.3$) positions). Finally, a difference was found between central and upper positions in low pitch, 75% Low pitch ($p<0.001$) and between down and up positions, low pitch, 75% Low pitch group ($p=0.006$). There was no difference between down and up positions in low pitch, 75% High pitch group ($p>0.6$).

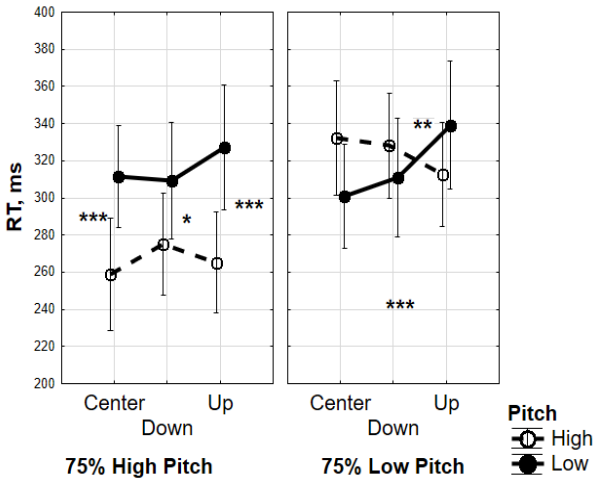

Figure 2: Experiment 2. A spatial position by pitch by pitch probability interaction. Vertical bars denote 95% CI. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

The pitch probability manipulation had a surprising and a strong effect: it made the pitch-position association to disappear. Note that Experiment 1 showed the association effect only in the upper position. The 3-way interaction (cf. Figure 2) unambiguously indicated complete disappearance of the effect with high expectancy of low pitch (75% Low pitch), moreover, it did not show an effect of expectancy per position, neither in down or control/central positions. Unlike 75% Low pitch probability, the 75% High pitch expectancy showed an expectancy effect across all three spatial positions. In a way, the expectancy effect also made the pitch-position association to disappear: it seems clear that the pitch-upper position association effect is observed mainly due to pitch expectancy. Generally, the results suggest that the endogenous manipulation led to dramatic and asymmetric change of the pattern of the results observed in Experiment 1. The latter clearly speaks in favor of non-automaticity of the cross-modal correspondence effect. To determine at which temporal stage of information processing the cross-modal corresponding effect took place (Experiment 1) or disappeared (Experiment 2), time distribution analyses were run.

RT Distribution Analyses

RT distribution analyses were conducted on data of each experiment tracing temporal dynamics of the correspondence effect, testing whether it occurs at earlier RTs and then decays or is developed over time. The pattern of effect's behavior would provide information on the level of automaticity of the effect that is, if the effect is automatic and takes place at the perceptual level (Evans & Treisman, 2010; Evans, 2020), it should be observed early in time. In addition, the analyses would trace the effect of probability manipulation (hence, the endogenous processes) over time.

For more details and clarity, the data were not reduced into the effect size as it is often done (e.g., Proctor, Miles, &

Baroni, 2011) and the conditions were preserved. Individual participants' data were rank-ordered per pitch and spatial position and then divided into five bins of 20%. Mean RTs per subject and per condition were computed per bin. Within-subject rANOVAs with factors pitch (high vs. low), spatial position (center vs. down vs. up) and bins (1-5) were conducted per experiment. Obviously, some results of the analyses mirrored the results presented above. These were only reported and not commented. Since the key interest was the time distribution of the effect across bins and spatial positions, the 3-way interactions were submitted to Bonferroni post-hoc test independently of significance of the interactions. Graphical representation of the results is shown in Figure 3.

Bin analysis of Experiment 1 showed a significant main effect of position ($F(2,70)=4.04$, $p=0.022$, $\eta_p^2=0.10$), main effect of pitch ($F(1,35)=4.51$, $p=0.040$, $\eta_p^2=0.11$), and an interaction between pitch and position ($F(2,70)=9.74$, $p<0.001$, $\eta_p^2=0.22$). Naturally, a significant main effect of bins ($F(1,140)=333.92$, $p<0.001$, $\eta_p^2=0.91$) was also found. Post-hoc analysis showed that significant differences were present across all bins (all ps<0.001). Pitch by bins, pitch by position, and pitch by position by bins interactions were not significant (all ps>0.2). The post-hoc test revealed only one significant difference between the pitches that was in upper position, fifth bin ($p=0.04$) (cf. Figure 3). Thus, our data did not provide strong evidence in favor of automaticity of the effect. The difference was evident only in the upper position and only in the last bin, even though there were visible numerical differences starting from bin 1.

Analyses of two probability groups were run separately. rANOVA for bin data of 75% Low pitch group obtained no main effects of either pitch (p>0.2) or position (p>0.06), and a significant two-way interaction between position and pitch ($F(2,68)=15.89$, $p<0.001$, $\eta_p^2=0.32$). A significant main effect of bins ($F(4,136)=277.83$, $p<0.001$, $\eta_p^2=0.89$) with significant differences between all bins (all ps<0.001) was also found. Further, a significant two-way interaction was found between pitch and bins ($F(4,136)=4.60$, $p=0.002$, $\eta_p^2=0.12$). Position by bins interaction was not significant (p>0.8). Finally, the three-way interaction was obtained ($F(8,272)=4.75$, $p<0.001$, $\eta_p^2=0.12$). Significant differences between pitches were found only in bin 5, upper position (p<0.001), and in bin 4, center position ($p=0.035$, see Figure 3). In agreement with the previous analysis (cf. Figure 2) the results suggested that the probability manipulation did not influence the lower position at all. Importantly, the conflict in upper position was observed in the 5th bin in spite of the manipulation. Practically, the 75% Low pitch manipulation did not change the (absence of) dynamics of the cross-modal correspondence pattern in comparison to the bin analysis of Experiment 1.

Bin analysis for 75% High pitch group showed a significant main effects of position ($F(2,70)=3.40$, $p=0.039$, $\eta_p^2=0.09$) and pitch ($F(1,35)=50.77$, $p< 0.001$, $\eta_p^2=0.59$). Significant main effect of bins ($F(1,140)=135.20$, $p=0.001$,

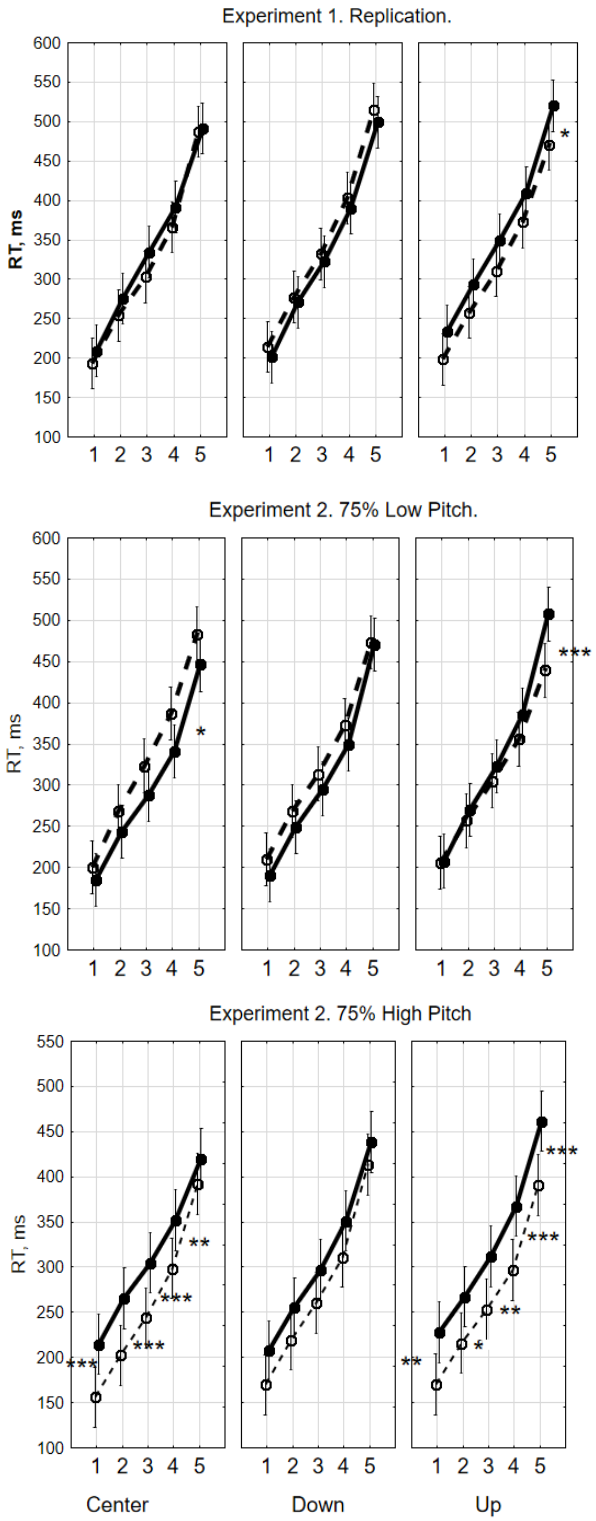

Figure 3: RT distribution across 5 bins, Experiment 1 (replication) and Experiment 2 (pitch probability manipulation). Dashed lines represent high pitch, solid lines – low pitch. Horizontal line represents bins from 1 to 5. Vertical bars denote 95% CI.
* p<0.05; ** p<0.01; *** p<0.001.

η_p^2=0.79) showed significant difference across all bins (all ps<0.001).

No other significant main effects or interactions were found (all ps>0.1). Significant differences between pitches per bin and position (sf. Figure 3) were as follows: center position, from bin 1 to 3 (p<0.001) and bin 4 (p=0.004), and upper position, in bin 1 (p=0.003), bin 2 (p=0.025), bin 3 (p=0.004), and bins 4 and 5 (p<0.001). Other differences were insignificant (all ps>0.3). Here the probability manipulation influenced both control and upper position but then, again, the lower position remained practically indifferent to the manipulation. Note that the influence of endogenous process (pitch expectation) showed itself very early in time, for control/central and upper positions.

Taken together, the distribution analyses confirmed and extended the results of Experiments 1 and 2, suggesting that the cross-modal association effect is prone to endogenous process influence, and is asymmetric in regard to both pitch and spatial position manipulations.

Conclusion

In the last couple of decades, processes behind audiovisual cross-modal correspondences have been a topic of an ongoing debate in the literature (Spence & Deroy, 2013). One of the main questions in the field is whether the cross-modal correspondence occurs in an automatic manner or whether an endogenous component is or can be involved.

While some studies have suggested that cross-modal correspondences occur in an automatic manner (Evans & Treisman, 2010; Evans, 2020), others showed a possible top-down process involvement behind the cross-modal correspondence (Chiou & Rich, 2012; Janyan et al., 2022). Nevertheless, these studies use rather diverse methodologies, which has led to a disparity in the research on the topic. The current study approached the problem differently by introducing top-down process as an expectation of a particular task-related stimulus. The goal was to test whether bottom-up and top-down processes would interact within the cross-modal paradigm or not. The interaction or modulation of the cross-modal effect by endogenous process would speak in favor of non-automaticity disagreeing with earlier studies (Evans & Treisman, 2010; Evans, 2020). Furthermore, it would provide evidence of endogenous element involvement into the cross-modal correspondence effect and would be in accord with the theory of Tang et al. (2016) on interaction of endogenous and exogenous processes and, importantly, on 'fast' endogenous processes that can 'act' in an exogenous manner.

Two experiments were conducted using a modified classical pitch height by object position in vertical space paradigm (Evans & Treisman, 2010). The modifications were the following: (i) a control condition was introduced (central object position); (ii) the standard 'congruency' factor was unfolded into more detailed pitch by position design. Experiment 1 replicated a cross-modal correspondence effect suggesting, however, that the effect came solely from the upper position, similarly to the results obtained by Janyan et al. (2022). Experiment 2 manipulated pitch probability between-participants, using the same paradigm as in Experiment 1. The results of both experiments showed that the pitch-position effect is asymmetrical, with lower position being impenetrable even by high probability of low pitch. Time distribution analyses confirmed the 'special' state of lower position across five bins. These results can be explained by general vertical asymmetry in visual processing (Previc, 1990). Specifically, lower visual field has stronger connections with the dorsal pathway, which is linked with visually guided actions, while the upper visual field is more connected to the ventral pathway, which is associated with perceptual identification of objects (Goodale & Milner, 1992). In addition, as it was mentioned before, it can be interpreted by natural statistics of visual and auditory scenes that show strong correlation between high frequency sounds and upper spatial position (Paris et al., 2014) and not that strong one between low pitch and lower spatial position. Thus, in the upper space the conflict between low pitch and the space is much stronger and, therefore, observable.

Our main focus, however, was on the endogenous influence on the cross-modal correspondence effect. The results of pitch probability manipulation suggested a not clear-cut interpretation. On the one hand, the pitch probability manipulation obviously influenced the effect, practically eliminating it. This can be taken as an indication of a non-automaticity of the correspondence effect (Moors & De Houwer, 2006) since the effect was impeded by the simultaneous additional information load (expectation). Thus, the results disagree with the suggestion and earlier data on automaticity of the cross-modal effect and its independence of selective attention (Evans & Treisman, 2010; Evans, 2020). In addition, another criterion of automaticity was also violated – that of speed (Moors & De Houwer, 2006). Time distribution analyses showed a significant effect only in the fifth bin, for both Experiment 1 (upper position) and Experiment 2 (upper position, low pitch expectancy). Thus, the effect developed slowly in time hence, it did not happen at the perceptual level or, at least, was not strong enough at the perceptual level.

On the other hand, the results showed a contrasting behavior of pitch height and spatial positions. For instance, while highly expected high pitch showed expectation effect in central and upper spatial positions, it did not show it in the lower position. And then, surprisingly, the low pitch expectancy practically did not manifest itself in neither position, even across the bins. These results obviously require further studies on vertical asymmetry focusing on different visuo-auditory perceptual, top-down, and attentional processes in regard to the cross-modal correspondence effect.

All in all, our results are in favor of non-automaticity of the cross-modal correspondence effect and in favor of the account proposed by Tang et al (2016) on interaction of endogenous and exogenous processes in multisensory integration.

Acknowledgments

We are grateful to our anonymous reviewers for their useful thoughts, comments, and suggestions. Unfortunately, we could not address all of them due to space requirements.

References

Chiou R., Rich A. N. (2012). Cross-modality correspondence between pitch and spatial location modulates attentional orienting. *Perception*, 41, 339–353.

Evans, K. K. (2020). The role of selective attention in cross-modal interactions between auditory and visual features. *Cognition*, *196*, 104119.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of vision*, *10*(1), 6-6.

Getz, L. & Kubovy, M. (2018). Questioning the automaticity of audiovisual correspondences. *Cognition*. 175. 101-108.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, *15*(1), 20-25.

Greene, H. H., Brown, J. M., & Strauss, G. P. (2019). Shorter fixation durations for up-directed saccades during saccadic exploration: A meta-analysis. *Journal of Eye Movement Research*, *12*(8).

Janyan, A., Shtyrov, Y., Andriushchenko, E., Blinova, E., & Shcherbakova, O. (2022). Look and ye shall hear: Selective auditory attention modulates the audiovisual correspondence effect. *i-Perception*, *13*(3), 20416695221095884.

Moors A., De Houwer J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, 132(2), 297–326.

Parise, C. V., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test. Experimental Brain Research, 220, 319-333.

Parise, C. V., Knorre, K., & Ernst, M. O. (2014). Natural auditory scene statistics shapes human spatial hearing. Proceedings of the National Academy of Sciences, 111(16), 6104-6108.

Previc, F. H. (1990). Functional specialization in the lower and upper visual fields in humans: Its ecological origins and neurophysiological implications. Behavioral and Brain Sciences, 13(3), 519-542.

Proctor, R. W., Miles, J. D., & Baroni, G. (2011). Reaction time distribution analysis of spatial correspondence effects. *Psychonomic Bulletin & Review*, *18*, 242-266.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-prime, Version 1.1. *Pittsburgh, PA: Psychology Software Tools*.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*, 971-995.

Spence, C., & Deroy, O. (2013). How automatic are crossmodal correspondences? *Consciousness and Cognition*, *22*(1), 245-260.

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences*, *14*(9), 400-410.

Tang, X., Wu, J., & Shen, Y. (2016). The interactions of multisensory integration with endogenous and exogenous attention. *Neuroscience & Biobehavioral Reviews*, *61*, 208-224.