이학석사 학위논문

The Aspect of Beat Perception in Audiovisual Synchronization

시청각 동기화 상황에서의 박자감 지각의 양상

2021 년 2 월

서울대학교 대학원

협동과정 인지과학

최 예 슬

The Aspect of Beat Perception in Audiovisual Synchronization

시청각 동기화 상황에서의 박자감 지각의 양상

지도교수 이 경 민

이 논문을 이학석사 학위논문으로 제출함

2021 년 2 월

서울대학교 대학원

협동과정 인지과학

최 예 슬

최예슬의 이학석사 학위논문을 인준함

2021 년 2 월

위 원 장 김 청 택 (인)

부 위 원 장 이 경 민 (인)

위 원 고 성 룡 (인)

Abstract

This study focuses on the association between visual and auditory stimuli and the effects of motor control that allows us to present two different sensory stimuli at the same time. We observed that the deviation of predictive response(DPR) is different between two condition: One is controlling the timing of visual stimuli with given periodic auditory stimuli so that two different sensory stimuli can be synchronized(A-V Condition), and the other is manipulating the timing of the auditory stimulation according to the presented periodic visual stimulus(V-A Condition). DPR is much more accurate and less variable in A-V condition compared to V-A condition, and this accuracy of DPR in each condition is dependent on IOIs(Inter-onset Intervals) of Stimuli given as a cue. This might suggest that the role of action control in the auditory domain is different from the visual domain as a form of beat perception(BP), and accuracy of predictive response for each sensory domain is up to IOIs of given stimuli.

Keyword: audiovisual synchrony, motor control, multisensory integration, time prediction, beat perception

Student Number: 2017-25502

Table of Contents

Contents of Figures and Tables

1. Introduction

Previous studies indicate that predicting the timing of upcoming beats, known as “phase alignment“, is human-specific tendency(Patel & Iverson, 2014). For example, human can align their taps just in time with metronome clicks spontaneously and it is widely known as the result of predictive behavior which anticipates the interval of the metronomic beats. Under this concept of Beat Perception and Synchronization(BPS), a number of studies focus on periodically repeated auditory stimulation so far. Patel & Iverson(2014) reported when auditory stimulus repeated in a specific pattern occurs, the listener tap one’s head or limbs to minimize the temporal difference between one’s movements and given beats.

Although Jones’s “Dynamic Attending Theory“(Jones and Boltz, 1989) indicates that temporal prediction of beats can be accurate even without these movements, Grahn and Rowe(2009) also found that the prediction of putative beats is engaged in putamen, supplementary motor area(SMA), and premotor cortex(PMC), that consist of a cortico-subcortical network. Jäncke et al(2000) also reported that SMA shows significant activation in audio-motor synchronization unlike visuo-motor synchronization. And above all, putamen is known as the core structure of beat and rhythm processing(Coull et al., 2011, Grahn and Rowe, 2009, Kotz et al., 2009, Teki et al., 2011, Wiener et al., 2009). Study by Grahn et al(2011) compared beat perception between two sensory sequences and observed that the beat perception was

more sensitive and showed higher putamen activation in auditory stimuli compared to visual stimuli. In addition, 38 meta-analysis studies using finger tapping tasks, audio-motor synchronization continuously activated putamen of basal ganglia, but visuo-motor synchronization using visual stimuli did not(Witt et al., 2008). Lots of findings consistently suggest that the role for the motor system, specifically the premotor cortex, is associated with the timing in auditory sequences and predicting its structure.

Then, the question arises from this: will the accuracy of the predicted response to the periodically repeated visual stimulus be different from that observed in the response to the auditory stimulus? In fact, visuo-motor synchronization has been reported to be dominant in processing spatial properties rather than time perception, like processing moving stimuli(Hove and Keller, 2010, Hove et al., 2010, Iversen et al., submitted for publication). It might not be that surprising, since it is generally known that vision excels at spatial resolution rather than temporal resolution. Shelton and Kumar(2010) found that auditory stimulation reaches the motor cortex faster than visual stimulation in this context. More specifically, it was reported that the time for general auditory stimulation to reach the brain after being received by the sensory device is about 8 to 12 ms, compared to 20 to 40 ms in the case of visual stimulation. It has been also observed that the direction and influence of motor control to visual and auditory stimulation are different from each other in the results of actual behavioral experiments as well as neural mechanisms in the brain. For example, auditory-motor interaction such as producing tapping performance to the beat occurs in the PMC and SMA(Grahn and Rowe,.

2009), while the visual-motor interaction is engaged in occipital lobe and parietal lobe of the brain(Culham et al., 2006). So numerous studies mainly focus on this difference of neural mechanism, and show similar results of behavioral experiment.

Therefore, BPS might not be an important property that visual system has to deal with. In studies of sensorimotor synchronization mainly using finger tapping, asynchronies tend to be greater in synchronization with isochronous visual sequences which is often given as light flashes, than in auditory sequences, usually composed of clicks or tones(Bartlett & Bartlett, 1959; Dunlap, 1910; Fraisse, 1948; Klemmer, 1967; Kolers & Brewster, 1985; Repp & Penel, 2002; Repp, 2003). Patel et al(2005) had also mentioned that the same rhythmic patterns can give rise to a clear sense of a beat when presented as sequences of tones but not when presented as sequences of flashing lights. In the same context, finger tapping is much more accurate with auditory stimuli than with flashing visual stimuli(e.g., Chen et al., 2002, Dunlap, 1910, Kolers and Brewster, 1985) and shows stable synchronization at much faster rates with auditory sequences(Repp, 2003).

It found that the participant has difficulty when the temporal gap between the stimuli, IOI(Inter-onset intervals), is 200-250ms or less in auditory condition and below 450-500ms in visual condition(Repp, 2003). It also indicated tapping with rhythmic sequences can be successful at least 400ms IOI for visual stimuli, and 150-200ms for auditory stimuli. Similarly, 400-800ms IOI is known to produce reliable beat perception so that this range of rates are optimal for synchronized tapping(Ono et

al., 2005; Drake et al., 2000a; McAuley et al., 2006). London(2012) also suggested humans perceive beats in a range of about 250ms to 2000ms, and especially, 400 and 1200ms give rise to the clear sense of beat. It can be inferred that the perceptible threshold of IOI is lower in auditory than in visual sequences, and temporal prediction of the auditory stimulus is more accurate than that of the visual stimulus even in a much shorter interval.

Based on this, Patel et al(2005) constructed periodic auditory and visual sequences in which IOI is constantly presented as 200ms/400ms/800ms, and measured BPS of each isochronous sensory stimuli. Like previous studies, the accuracy of response was significantly less in the visual condition than in the auditory condition, Also, Repp and Penel(2004) observed that the prediction for auditory sequences is more accurate than visual, when IOI is less than 2000ms. As such, most of the BPS-related studies has reported similar result that even if both are isochronous sequences, temporal prediction and figger tapping performance for auditory stimuli is more accurate and successful than visual stimuli.

But not considering multisensory integration in our daily life might be the blind spot of all these previous studies. For example, we don't simply move our fingers in response to sound when we play music. We visually perceive the temporal representation of the notes on the sheet music, then move our hands to play the instrument. So this study aims to verify the reliability of the previous studies and create an audiovisual integration situation. In this study, the auditory stimulus has

to be produced through finger tapping simultaneously according to the isochronous visual cue, or conversely, manipulating visual stimulus according to the auditory cue so that two sensory sequences be generated simultaneously. Through this, we will observe how the prediction of periodicity in the auditory-motor and visual-motor interactions changes when they are connected with the other sensory stimuli. Deviation of predictive timing(DPR) is measured from the response of participants, from which we can judge whether the prediction of the sequences is accurate or not. From this experiment, we would see how temporal prediction for each sensory sequences occurs in a situation of multisensory integration.

2. Method.

2.1. Sync-Action Task

Experimental stimuli were presented to a Windows-based computer using a psychtoolbox-based Matlab program, and a 17" LCD monitor and earphones were used to present the auditory cue, which was presented as a beep sound. The visual cue was shown as a black circle that flashes periodically at the center of the screen. Both auditory and visual cues were presented at regular intervals.

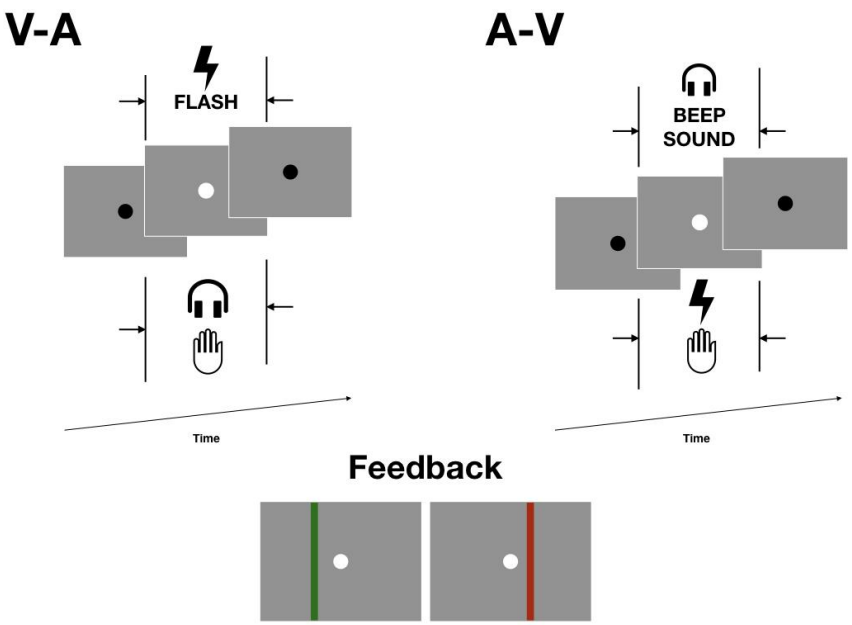


Figure 2 Sync-Action Task

This experiment contains two conditions; In V-A condition, visual cue is presented in flash and a participant should press the keyboard with the finger in order to produce beep sound which he or she hears with earphones so that the auditory sequences produced by participant must be synchronized with given visual sequences. In contrary, the participant has to manipulate the flash simultaneously when the beep sound is presented in A-V condition. For the intervals of the stimuli, this study referred to previous studies suggesting optimal range of IOIs. Based on these several findings, each condition of this study consists of periodic auditory or visual cue with constant IOI of 500ms, which is quite fast rate to perceive, 800ms as a moderate and maybe optimal rate, 2000ms as the limit rate that we can barely perceive.

In all conditions, feedback is provided on the screen to assist participants' task performance. If the participant manipulates the sensory stimulus earlier than the presented cue, a red bar is displayed, and if the participant operates it more slowly, a green bar is displayed on the screen. As the participant succeed in synchronizing a timing of the stimulus with the onset of the presented cue, the thickness of the rod becomes thinner.

Data of participants' response are calculated and recorded as DPR, abbreviated for 'Deviation of Predictive timing' in this study. It represents the deviation between the time at which the stimulus was presented and the predicted response to it. DPR is recorded along with the number of times the participant pressed the keyboard. If the subject responds earlier than the presented cue, DPR is expressed in

the form of '-', and if the subject responds late, it is expressed as a value of '+'.

2.2. Participants

Participants were recruited through the portal site of Seoul National University, SNULife. The experiment consists of 25 young healthy adults(n = 25, mean age = 23.91667, ± 5.468228).

2.3 Procedure

To perform the task, Participant needs to look at the computer screen at a distance of 30cm from the monitor and press the keyboard to operate the timing of the stimulus. In response to the auditory or visual cue that appears periodically, the participant presses the keyboard and manipulates corresponding sensory stimulus to be temporally synchronized with the cue. When the experiment begins, the participant operates the keyboard so that the manipulated stimulus is simultaneously presented from the presented cue at regular intervals. There are three intervals: 500ms, 800ms, and 2000ms. The A-V condition and V-A condition including these three types of intervals are presented randomly. Therefore, participant performs a set of experiments consisting of six conditions. Each condition has 40 trials.

3. Result

3.1. General Result

First, this study tested whether DPR of each participant follows a normal distribution or not. The Shapiro-Wilk normality test was used to test whether the data follow a normal distribution. The results using Shapiro-Wilk test and descriptive statistics of the data are as follows(see **Table 1**, **Table 2**, **Table 3** and **Table 4**).

Table 1 P-value of each participant's DPR in A-V condition using Shapiro-Wilk test for normality

		Inter-Onset Interval (IOI)		
		500ms	800ms	2000ms
Participant ID	1	3.693e-07***	1.277e-07***	1.601e-11***
	2	1.24e-07***	0.004759**	6.503e-10***
	3	0.007704**	2.795e-09***	4.247e-13***
	4	0.001756**	4.923e-05***	3.358e-05***
	5	3.252e-05***	**0.1686**	8.37e-05***
	6	2.077e-08***	0.006187**	1.205e-07***
	7	5.086e-06***	0.003818**	< 2.2e-16***
	8	2.156e-10***	0.007089**	6.176e-05***
	9	0.0001271***	0.01701*	9.613e-09***
	10	1.617e-07***	0.01391**	0.0003543***
	11	2.267e-12***	**0.2467**	4.953e-05***
	12	0.001103***	2.108e-07***	0.002346**
	13	3.966e-05***	0.004369**	1.944e-05***
	14	**0.1441**	0.03588*	4.528e-11***
	15	8.041e-09***	0.03293*	6.996e-09***

16	2.954e-09***	**0.221**	1.254e-10***
17	0.009195**	0.007642**	2.647e-08***
18	1.379e-08***	**0.05816**	**0.04542**
19	5.913e-10***	0.01702*	1.421e-09***
20	0.002735**	**0.07844**	5.239e-09***
21	0.004521**	**0.2741**	5.756e-11***
22	1.691e-05***	2.474e-10***	1.066e-09***
23	**0.05174**	**0.3262**	0.01011*
24	1.863e-09***	**0.005184**	1.598e-07***
25	2.966e-12***	6.91e-07***	0.0003611***

Table 2 P-value of each participant's DPR in V-A condition using Shapiro-Wilk test for normality

		Inter-Onset Interval (IOI)		
		500ms	800ms	2000ms
Participant ID	1	7.092e-07***	1.47e-09***	1.159e-12***
	2	7.372e-06***	2.63e-09***	0.2528
	3	0.06407	0.02083*	4.44e-09***
	4	0.0002111***	0.05497	0.0008965***
	5	0.01876*	0.127	0.0002089***
	6	0.0863	0.01829*	0.01398**
	7	0.0313*	4.943e-07***	5.645e-09***
	8	1.025e-07***	1.025e-07***	4.218e-05***
	9	0.004315**	0.004315**	2.513e-08***
	10	0.7917	0.0251	8.779e-08***
	11	5.804e-07***	7.681e-12***	0.0004131***
	12	2.698e-06***	0.3165	2.463e-05***
	13	0.4192	2.942e-12***	0.0004439***
	14	0.3302	0.04032*	0.0206*
	15	0.1043	0.2609	0.001458**
	16	3.419e-07***	0.2683	2.905e-05***
	17	0.007476**	0.1469	8.321e-08***
	18	0.0235*	0.02999*	3.555e-08***
	19	3.542e-09***	0.001348**	0.078
	20	0.03622*	0.006343**	9.468e-07***
	21	0.001806**	0.003252**	1.062e-13***
	22	0.6199	0.02977*	2.793e-10***
	23	0.009759**	0.2413	0.03084*
	24	0.2782	0.2003	0.2673
	25	5.642e-07***	0.001586**	5.406e-09***

Table 3 Central tendency and dispersion of DPR in A-V condition

		Inter-Onset Interval (IOI)		
		500ms	800ms	2000ms
Central Tendency	mean	0.005732262	-0.005913311	0.03261137
	median	-0.0114449	-0.01658875	0.0178984
Dispersion	sd	0.09967941	0.1211727	0.2775618
	var	0.009935986	0.01468282	0.07704053
	IQR	0.0750549	0.1006198	0.3530562
	range	-0.4181817 0.4190109	-0.5839962 0.5992984	-1.018045 1.873119

Table 4 Central tendency and dispersion of DPR in V-A condition

		Inter-Onset Interval (IOI)		
		500ms	800ms	2000ms
Central Tendency	mean	0.02860113	0.06910076	0.0560997
	median	0.02413	0.05447845	0.0877891
Dispersion	sd	0.1387179	0.140515	0.3296807
	var	0.01924266	0.01974447	0.1086894
	IQR	0.1350868	0.1527518	0.4135596
	range	-0.3642577 0.3740255	-0.5838975 1.1347746	-1.578387 1.713201

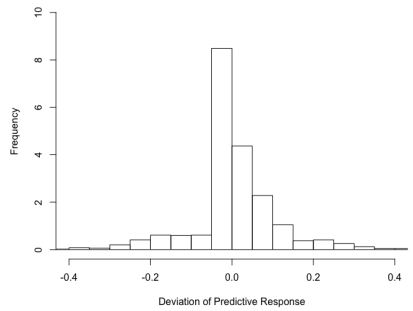
Though several participants show normal distribution in both A-V and V-A conditions in every IOIs, conducted analysis using same Shapiro-Wilk test shows that the data don't follow normal distribution in every conditions when all the participants are put together(see Table 5).

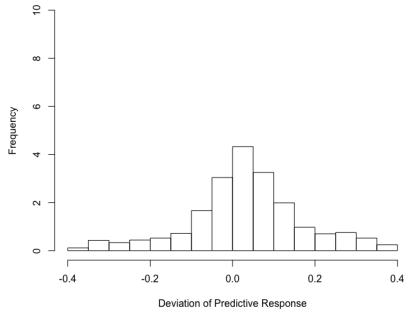
Table 5 P-value of 25 participants' DPR in each condition using Shapiro-Wilk test for normality

Condition	Inter-Onset Interval (IOI)	p-value
AV	500ms	< 2.2e-16***
	800ms	< 2.2e-16***
	2000ms	< 2.2e-16***`
VA	500ms	< 2.2e-16***
	800ms	< 2.2e-16***
	2000ms	< 2.2e-16***

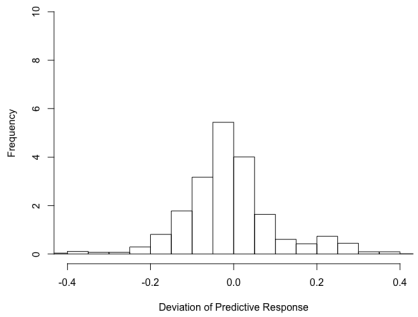
In every 500ms/800ms/2000ms IOIs, the p-values of both A-V Condition and V-A Condition including 25 participants' whole data of DPR were less than 2.2e-16, so we assume that DPR in all conditions of this experiment tend not to follow the normal distribution. This observation was confirmed in with the histogram below(See Table 5 and also Table 6).

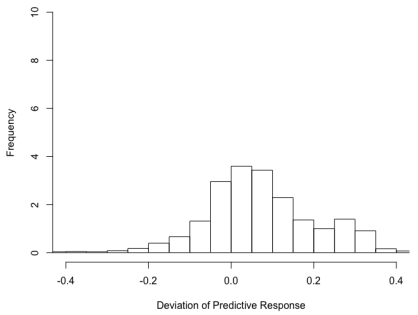
A-V Condition	V-A Condition



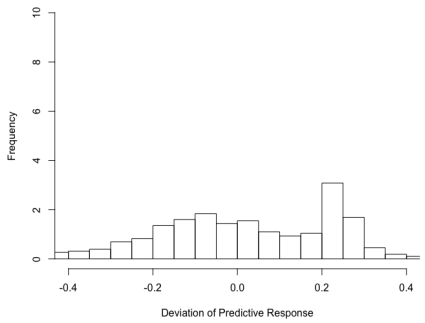


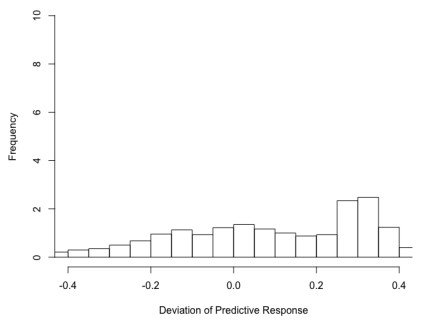
IOI 500ms





IOI 800ms





IOI 2000ms

Table 6 Histograms of Each Condition's DPR

If DPR does not follow a normal distribution, the difference between the A-V and V-A conditions and the difference between the IOI conditions can be analyzed with non-parametric techniques like Wilcoxon rank sum test and Kruskal-Wallis test. First, the difference between the two condition according to each IOIs was compared using the Wilcoxon test(Table 7), and then the difference between IOI 500ms/800ms/2000ms in each condition was tested by Kruskal-Wallis test(Table 8).

Table 7 The difference between the A-V and V-A conditions using Wilcoxon rank sum test

Inter-Onset Interval	p
500ms	4.459e-0.6***
800ms	< 2.2e-16***
2000ms	3.188e-10***

Table 8 The difference between the IOI conditions using Kruskal-Wallis test

Condition	p
A-V	1.304e-11***
V-A	<2.2e-16***

As a result of performing the Wilcoxon rank sum test, p-value

significantly small in all conditions(See Table 7). So in all three IOIs, DPR in A-V condition showed a significant group difference from DPR in V-A condition. And then Kruskal-Wallis test was done to observe the difference between 3 IOIs in each of the two conditions. We observed that the statistical difference according to the IOIs was also significant(See Table 8). In summary, DPR was significantly different according to the IOIs in both conditions.

Therefore, the DPR in the A-V condition is significantly different from the DPR in the V-A condition in all IOIs. And in both conditions, the difference according to IOIs was also significant.

3.2. Analysis of Result Using Kernel Density Estimation

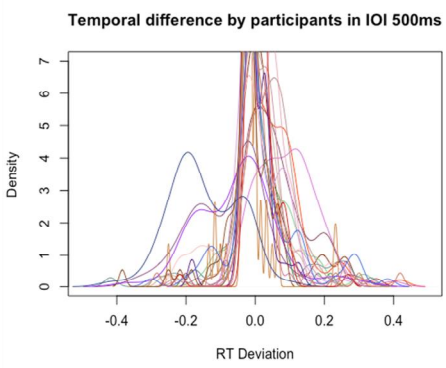
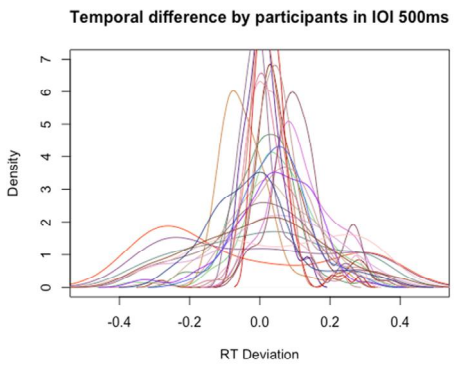
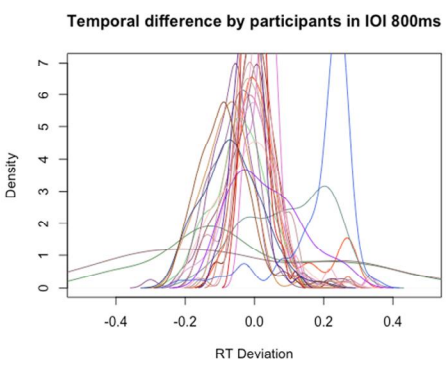
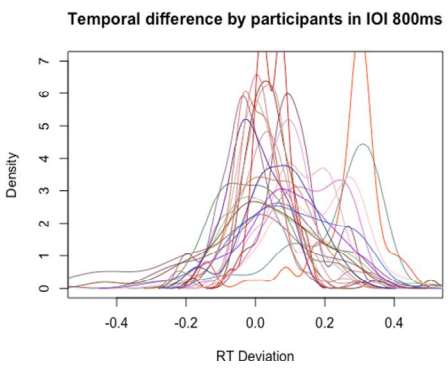
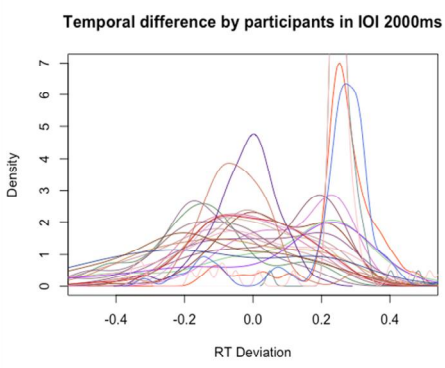
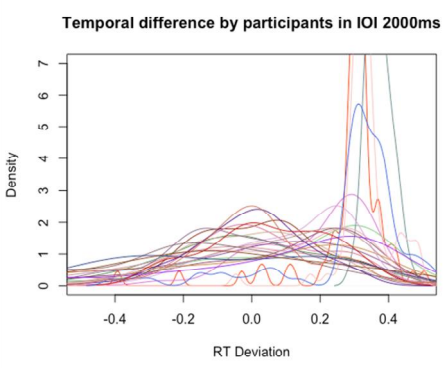
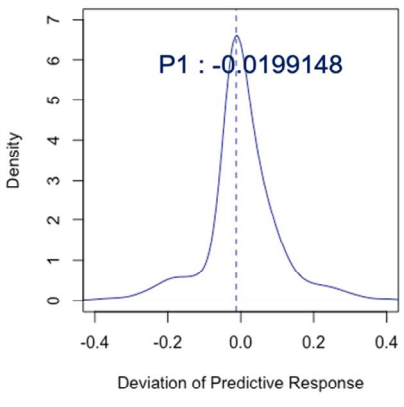
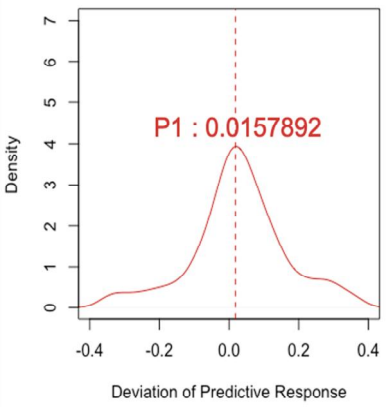
A-V condition	V-A condition
	
	
	

Table 9 Histogram distribution of every participant's DPR

A-V condition	V-A condition

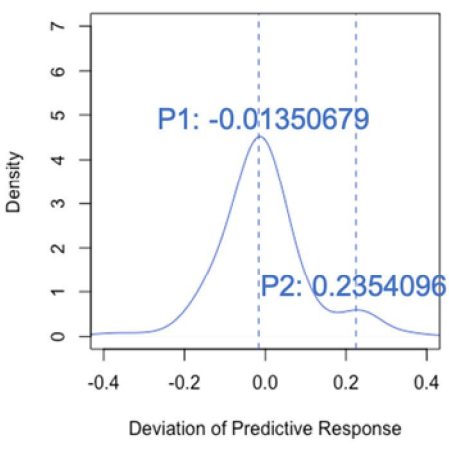


(a)

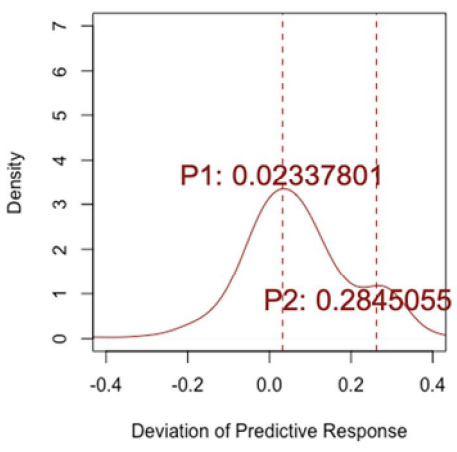


(b)

IOI 500ms

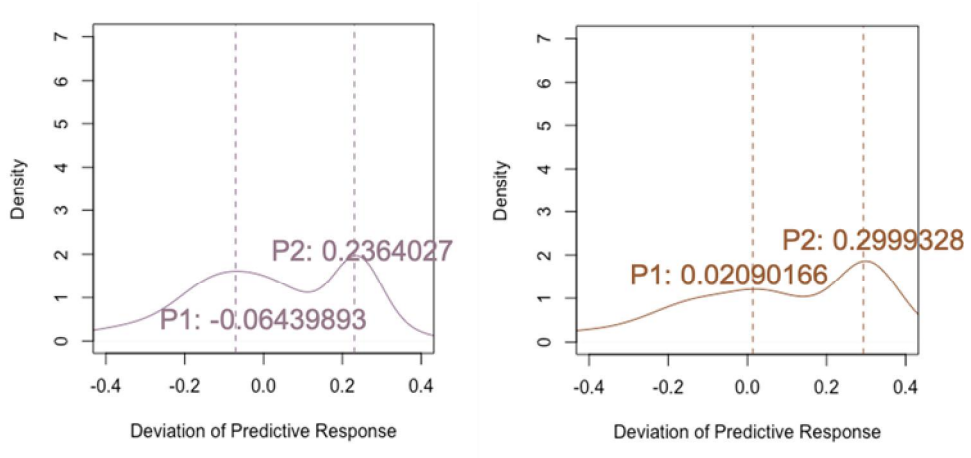


(c)



(d)

IOI 800ms



(e) (f)

IOI 2000ms

Table 10 DPR distribution of participants using Kernel Density Estimation

If all DPRs in each 6 condition are significantly different, then the next question is what these differences mean. To answer this, this study investigated the distribution of DPRs using Kernel density estimation(KDE) to find certain pattern of the data(Table 9 and Table 10).

Noticeably, bimodal distribution was observed in 800ms and 2000ms IOIs at both A-V and V-A condition, so two peaks are found in these conditions to be analyzed. Primary peak, which is also described as P1, is ahead of the secondary peak, P2(See Table 10). According to the graphs above, the temporal prediction does not randomly break down as the IOI increases, but appears in a bimodal pattern instead. P1 tends to converge to 0ms in 500-800ms IOIs, it decreases in size when IOI

increases. On the other hand, the secondary peak named P2 is observed constantly at around 200ms when IOI is over 800ms.

To analyze this bimodal distribution, the x value of these two peaks was calculated. P1 in A-V at 800ms IOI, which is (c) in **Table 10**, is around -13ms, which is the smallest absolute value when compared to (a) or (e). However, there is only a 7ms difference between (a) and (c), and is relatively not large compared to the 50ms difference between (c) and (e). In other words, P1 at 800ms is the closest to 0, but there was little difference between P1 at 500ms and 800ms IOIs. In all three IOIs, P1 was always recorded earlier than the timing of visual cue. In particular, P1 at 2000ms shows that it was recorded about 50ms earlier compared to the other two IOIs. The frequency, which is the y value of P1, tends to decrease as the IOI increases, but decreases significantly between 800ms and 2000ms.

In V-A condition, P1 of 500ms, (b) in **Table 10**, is about +15ms, which is the smallest absolute value among the three IOIs. At 800ms and 2000ms, P1 is at the point of +20ms(see (d) and (f) in **Table 10**). The difference between the former and the latter is about 5ms, which is not very large. In other words, P1 at 500ms has the smallest DPR, but difference of it according to IOIs may be insignificant. Unlike A-V condition, P1 was always delayed and the timing of auditory cue is prior to this. In the matter of frequency, P1 decreases as the IOI increases, and as in the A-V condition, it decreased particularly significantly between 800ms and 2000ms.

In Summary, DPR was the smallest in 500ms or 800ms, and the frequency of P1 decreased as IOI increased in both A-V and V-A condition. When comparing A-V and V-A condition, P1 in the A-V was about 35 ms earlier than in the V-A in both 500ms and 800ms. This difference increased to 85ms in 2000ms IOI, and P1 in the V-A showed a tendency to be delayed at all IOIs than in the A-V condition.

When comes the case of P2, it is consistently observed at about +235ms(see (c) and (e) in **Table 10**) in the A-V condition, and the frequency increases with increasing IOI. Similar results are founded in V-A condition. In this case, however, P2 is distributed around +300ms. When comparing the x value of P2 of A-V and V-A under 800ms IOI, the V-A shows P2 delayed by 50ms, and this delay slightly increases to 65ms under 2000ms IOI.

Taken together, P2 begins to emerge from 800ms IOI in both A-V and V-A conditions, and its frequency is observed to be greater at 2000ms IOI. Unlike P1, the x value of P2 shows a relatively small fluctuation depending on the IOIs, which is distributed at a certain point in each two sensory condition, and the frequency is directly proportional as the IOI increases.

4. Discussion

The smaller the DPR is, the more accurate the predictive response is. Therefore, P1, when the DPR distributed near 0 in every conditions, can be assumed to be the peak of the accurate predictive response. Conversely, P2, which appears as a point consistently delayed from the cue's timing, can be assumed to be just reactive to the given cue, so it can be regarded as a peak of reactive response.

In the A-V condition, P1 of 500ms and 800ms IOIs shows DPR distributed around -20ms and -13ms respectively, so it can be regarded as more accurate predictive response compared to P1 of 2000ms IOI which was recorded as -64ms. This means that predictive timing in the A-V condition is more accurate in 500ms or 800ms IOIs, rather than in 2000ms IOI.

In contrast, the x value of P1 in the V-A condition does not move significantly even when the IOI increases. This means that the accuracy of predictive timing in V-A is relatively less affected by IOI than in A-V condition. However, this does not mean that the predictive timing of V-A is more accurate than that of A-V. Except for 500ms IOI, DPRs of the A-V are always smaller than that of the V-A condition, which means A-V condition shows more accurate predictive response than V-A. Likewise, the frequency of P1 in the A-V in all IOIs is higher than in the V-A, which means that temporal prediction in A-V is more precise than that in the V-A under the same IOI. However, in both

conditions, we can observe that the frequency of P1 decreases with increasing IOI. It can be inferred that when IOI increases, the precision of predictive timing decreases regardless of cue's sensory module.

In the case of P2, which reflects the reactive response, it begins to appear at 800ms and the frequency of it is observed to be greater at 2000ms. We can think that the predictive response to the presented cue is activated in all IOIs, but the reactive response is hardly seen in 500ms IOI. This may be because 500ms IOI belongs to the most optimal range for performing auditory-visual synchronization as described above(Ono et al., 2005) and Bartlett and Bartlett(1959) once reported this as limits of synchronization.

Though the variability of asynchronies and inter-tap interval(ITI) increases with ITI duration, the problem is that there may not be enough time to use feedback when IOI is 500ms. This implies the possibility that the task performance at 500ms IOI might belongs to the system with open-loop control. In other words, predictive response may be generated while the performance is controlled by a series of rhythmic processes structured in advance, instead of generating a predictive response after performing a reactive response and modifying the response with feedback information about the cue. Over 800ms IOI, however, the response might be controlled by a closed-loop system, accepting auditory or visual information provided as cue, controlling it with feedback, creating a reactive response, and then correcting the error in the response. Although the IOI increases, P1 still remains at a small rate and it might indicate that the predictive response is

continuously generated in small amounts by this closed-loop control.

In the case of P2, which represents the reactive response, it is delayed by 235ms in both 800ms and 2000ms IOIs of A-V Condition, and distributed around 300ms in V-A. That is, as in predictive response, reactive response in V-A tends to be more delayed than in A-V. As the IOI increases, the frequency of reactive responses increases under both conditions. This might mean that P2 is likely a reactive response to form a closed-loop control in response to the cue itself when temporal prediction fails.

Shelton & Kumar(2010) once reported that the auditory reaction time is faster than the visual reaction time and the mean of the visual reaction time is around 331ms as compared to the mean of the auditory reaction time of around 284ms. Almost same result was observed in this study. In A-V condition when the cue is given as auditory stimulus, the mean of each P2 of 800ms and 2000ms IOI is around 236ms, and 292ms in the V-A condition. The difference between the auditory and the visual reaction time in the previous study was 47ms, and 56ms in this experiment which number is quite similar to each other.

This supports not only the previous studies, but also a new aspect of BPS in situation of multisensory integration. Like response to the auditory stimulus is faster than the condition using visual stimulus, The reactive response in the A-V condition, which binds auditory cue and visual stimuli together, also occurs faster than V-A condition. In

addition, compared to the cited study, each P2 in A-V Condition and V-A Condition is 50ms faster than audio and visual reaction times revealed in previous study. This may be because the motor-operated sensory stimulus synchronized with the external cue in each condition acts as a kind of feedback on task performance, which helps reducing the reaction time.

But still a question remains: it might be evident that the more IOI increases, the larger DPR's variance becomes. What if the relative phase of DPR is calculated, how would the results based on the above absolute phase change? Focusing on P1 and P2 in relative phase based on 2000ms IOI(**Table 11**), P1 in 500ms and 2000ms IOIs shows no big difference but P1 in 800ms IOI is found to be the most accurate DPR in A-V condition. So we can also suggest that optimal temporal window for auditory beat prediction might be 800ms IOI rather than 500ms IOI when considering relative phase.

Table 11 Relative phase based on 2000ms IOI of A-V and V-A condition

Condition	Inter-Onset Interval (IOI)	Peak	DPR
A-V	500ms	P1	-0.0796592
		P2	
	800ms	P1	-0.033766975
		P2	0.588524
	2000ms	P1	-0.06439893
		P2	0.2364027
V-A	500ms	P1	0.0631568
		P2	
	800ms	P1	0.058445025
		P2	0.71126375
	2000ms	P1	0.02090166
		P2	0.2999328

When it comes to V-A condition, P1 in 500ms and 800ms IOIs are not that different from each other but surprisingly less accurate than in 2000ms IOI(Table 11). P2 in 2000ms IOI is also more accurate than in 800ms IOI. According to this finding, prediction of visual rhythm is relatively better in 2000ms IOI than other IOIs considering the range of increased interval of visual cue. This unintuitive finding might also confirm the relative ignorance about the temporal prediction of the vision. In both A-V and V-A conditions, P2 in 2000ms IOI is still more accurate than 800ms IOI, which might means that reactive response is formed actively and even more accurately in 2000ms IOI.

So we can see the accuracy and precision of temporal prediction in A-V condition is consistently higher than that of V-A condition. It means that when external sensory cue is given in the form of auditory, temporal prediction is better than visual. As we saw, it is convincing because of the fact that audition excels at temporal prediction than vision. In other words, since given cue in the A-V condition is auditory stimulus, the accuracy and precision of beat perception and temporal prediction are superior than that of the V-A Condition in all IOIs. However, this predictive response is best achieved when IOI is given as 500ms and 800ms, and when it is 2000ms, it becomes difficult to perform successfully and the reactive response is facillitated as a counter effect.

If so, this accuracy and precision can be seen to be strongly dependent on the auditory characteristics of the temporally serial external auditory cue. When the external cue and the motor-driven stimulus are associated with each other as different sensory modules, the sensory module of the external cue is more likely to dominate than the motor-driven stimulus. More specifically, each temporal representation of two different sensory modules must be matched both internally and externally in this study, and due to dominance of the temporal representation of the exogenous cue, motor manipulating is entrained to this dominant sensory module. A diagram of this provisional model is shown below.

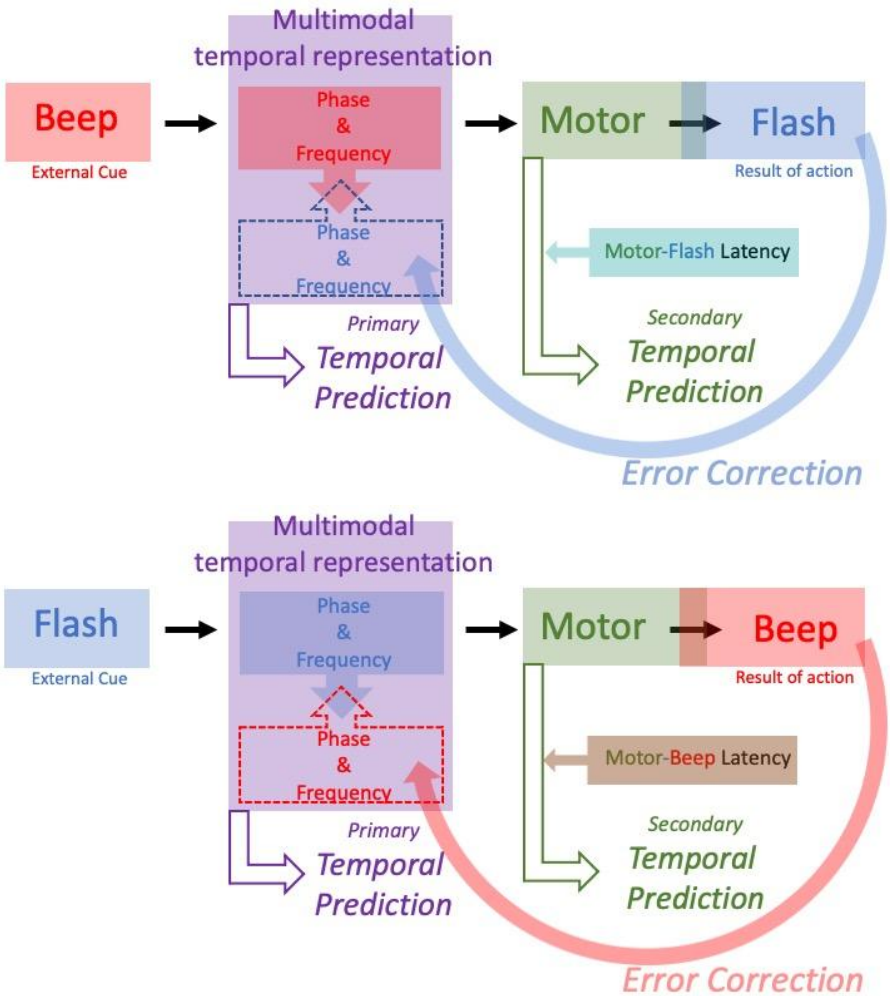


Figure 2 The Model of Temporal Prediction with Internal Temporal Representation in A-V and V-A Conditions

If temporal representation from auditory cue is created to make temporal prediction, it will take up a greater weight in producing movement than in the corresponding visual sequences which are temporarily generated, to be operated by motor. When these two temporal representations are combined together into multimodal

temporal representation, phase and frequency of external stimulus are used. The beat perception of each modality will not be neutralized even if the temporal representation of each modality is integrated, and the predictive response will be formed according to the temporal representation generated by the modality of the external cue.

This multimodal temporal representation is to form predicted sensory feedback named primary temporal prediction in Figure 2, which is distinguished from actual sensory feedback, written as secondary temporal prediction in the same figure. Primary temporal prediction occurs when the temporal representation of external cue is established so that it associates with internal temporal representations of which sensory module has to be manipulated later by motor. After multimodal temporal representation is delivered to motor command, secondary temporal prediction is made to estimate latency between motor and result of action. These temporal predictions are integrated to decide when action has to be made, and DPR from the result of action and feedback on screen act as error correction for more precise temporal prediction.

So P1, defined as predictive response in this study, is the result of this integrated temporal prediction. In 500ms IOI, however, error correction seems to be excluded because reaction response, which is P2 in this study, is rarely found. It is very interesting that DPR in this IOI condition is very small not having error correction even ITI in this condition seems too short to use feedback properly. It may be because this IOI is optimal range for accurate and precise temporal prediction

so that it does not need feedback to correct its predictive response. Thus it can be inferred that the temporal prediction in 500ms IOI might be controlled by open loop system as described above. Just two temporal predictions themselves are enough to predict IOIs of isochronous auditory sequences without error correction. Therefore, P1 in 500ms IOIs is composed of primary and secondary temporal prediction, which does not include error correction.

But the error correction seems to be necessary for temporal prediction over 800ms IOIs since the longer IOI becomes, the more difficult temporal prediction is. To predict the timing of the cue and respond to it simultaneously in 800ms and 2000ms IOIs, accuracy and precision of predictive response are decreased but reactive response increases in closed loop control system. In 800ms IOI, this reactive response which is marked as P2, is not that large compared to 2000ms IOI, and seems to act like a quite successful feedback for establishing appropriate temporal prediction. But this tendency breaks down when in 2000ms IOI, in which P2 is much larger than predictive response as P1. Error correction is not meaningful anymore because primary prediction starts to break up in predictable limit of IOI, which is given as 2000ms in this study(London, 2012).

So the reactive response P2, which occurs as the IOI increases, can be the part of the process that caused due to the formation of the inner temporal representation, and also can be regarded as a result of the motor operation that failed to form the proper temporal representation. In order to make temporal prediction about the IOI of the cue, reactive

reaction corresponding to the external cue is generated. But when the prediction fails, the frequency of these simple reactive motor reactions increases, resulting in P2. It is consistently delayed by 50 ms in the V-A than in the A-V regardless of IOIs. This shows that not only does the external visual cue form an internal temporal representation less successfully than the auditory cue, but the reactive motor response generated by the visual cue is also slower than the motor response formed by the auditory cue.

In summary, external cue with temporal information is accepted as an internal temporal representation and associated with the other internal temporal representation driven by motor. Then these two different sensory representations are combined to form a multimodal temporal representation. At this time, the internal temporal representation derived from the external cue is more dominant than the other representation in forming the temporal prediction. Motor-driven stimulus is generated as a result by manipulating the motor based on the temporal sequence of this multimodal representation. The result of action(beep or flash) executed by the motor will also function as a feedback that can correct errors in its internal temporal representation. For example, in the case of the A-V condition, it becomes difficult to predict the interval of the auditory cue as the IOI increases so that the internal time representation according to the cue can not be formed well. In 2000ms IOI, the temporal representation of the visual stimuli generated by the motor operation also can not be properly generated, so it may not be appropriately used as a feedback for predicting the periodicity of the auditory cue.

This study observed the BPS, which have been experimented with a single sensory module, in perspective of audio-visual integration mediated by motor control. Previous studies so far have focused on whether the response of BPS is simply 'more' accurate in which single sensory module. However, situations consist of two or more sensory modules, such as playing music or hitting a constant flying golf ball with an appropriate frequency in everyday life are not that simple.

In order to explore such phenomena more broadly, this study attempted to find an answer by presenting two multisensory conditions, A-V condition and V-A condition. It is meaningful that the distribution of response of BPS is specifically divided into predictive response and reactive response to examine which sensory module is 'how' rather than 'more' accurate at temporal prediction.

This study aims to provide an understanding of the process of multimodal integration in everyday life and some insight which can be applied to various fields in the 4th industrial revolution, like controlling action integrating sensory modules or speech process of Artificial Intelligence. In addition, this study provides insights into users' perception of avatars implemented in Virtual Reality for more natural movement and interaction in it, which aims to decrease a sense of incongruity but increase immersion of user playing it.

But there remains the question of 'why' the prediction of auditory sequences is still more accurate than the visual, just like reactive response to each sensory stimuli. What made human to predict the

frequency of sound better, not the flash? This is left as a task for further study to be investigated.

References

Bartlett, N. R., & Bartlett, S. C. (1959). Synchronization of a motor response with an anticipated sensory event. *Psychological review, 66*(4), 203.

Bertelson, P., & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception & psychophysics, 29*(6), 578-584.

Chen, Y., Repp, B. H., & Patel, A. D. (2002). Spectral decomposition of variability in synchronization and continuation tapping: Comparisons between auditory and visual pacing and feedback conditions. *Human movement science, 21*(4), 515-532.

Coull, J. T., Cheng, R. K., & Meck, W. H. (2011). Neuroanatomical and neurochemical substrates of timing. *Neuropsychopharmacology, 36*(1), 3-25.

Culham, J. C., Cavina-Pratesi, C., & Singhal, A. (2006). The role of parietal cortex in visuomotor control: what have we learned from neuroimaging?. *Neuropsychologia, 44*(13), 2668-2684.

Grahn, J. A., & Rowe, J. B. (2009). Feeling the beat: premotor and striatal interactions in musicians and nonmusicians during beat perception. *Journal of Neuroscience, 29*(23), 7540-7548.

Grahn, J. A., Henry, M. J., & McAuley, J. D. (2011). FMRI investigation of cross-modal interactions in beat perception: audition primes vision, but not vice versa. *Neuroimage, 54*(2), 1231-1243.

Hove, M. J., & Keller, P. E. (2010). Spatiotemporal relations and movement trajectories in visuomotor synchronization. *Music Perception, 28*(1), 15-26.

Jäncke, L., Loose, R., Lutz, K., Specht, K., & Shah, N. J. (2000). Cortical activations during paced finger-tapping applying visual and auditory pacing stimuli. *Cognitive Brain Research, 10*(1-2), 51-66.

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological review, 96*(3), 459.

Jose, S., & Gideon Praveen, K. (2010). Comparison between auditory and visual simple reaction times. *Neuroscience & Medicine, 2010.*

London, J. (2012). *Hearing in time: Psychological aspects of musical meter.* Oxford University Press.

Ono, K., Nakamura, A., & Maess, B. (2015). Keeping an eye on the conductor: neural correlates of visuo-motor synchronization and musical experience. *Frontiers in human neuroscience, 9*, 154.

Patel, A. D., Iversen, J. R., Chen, Y., & Repp, B. H. (2005). The

influence of metricality and modality on synchronization with a beat. *Experimental brain research, 163*(2), 226-238.

Patel, A. D., & Iversen, J. R. (2014). The evolutionary neuroscience of musical beat perception: the Action Simulation for Auditory Prediction (ASAP) hypothesis. *Frontiers in systems neuroscience, 8*, 57.

Repp, B. H. (2003). Rate limits in sensorimotor synchronization with auditory and visual sequences: The synchronization threshold and the benefits and costs of interval subdivision. *Journal of motor behavior, 35*(4), 355-370.

Repp, B. H., & Penel, A. (2004). Rhythmic movement is attracted more strongly to auditory than to visual rhythms. *Psychological research, 68*(4), 252-270.

Witt, S. T., Laird, A. R., & Meyerand, M. E. (2008). Functional neuroimaging correlates of finger-tapping task variations: an ALE meta-analysis. *Neuroimage, 42*(1), 343-356.

시청각 동기화 상황에서의 박자감 지각의 양상

최 예 슬
서울대학교 대학원
협동과정 인지과학 전공

박자감 지각은 음악을 들으면서 고개 혹은 손과 발을 까닥거리면서 박자를 맞출 때 주로 일어나는 것으로 알려져 있다. Patel과 Iverson(2014)에 의하면 주기성을 지닌 청각자극에 대한 예측의 시간적 정확성은 인간에게서 두드러지게 관찰되는 것으로 알려져 있다. 주로 BPS(beat perception and synchronization)에 대한 보고를 중심으로 이루어지고 있는 박자감 지각에 관한 연구들은 주기적으로 자극이 제시될 때 시각자극보다 청각자극에 대한 시간적 예측이 짧은 주기에서도 더 정확하다는 점을 일관적으로 보이고 있다(Repp, 2003; Repp, 2004; Patel et al, 2005). 이러한 선행연구를 바탕으로 본 과제는 시청각 통합 상황에서 시각자극에 반응하여 청각자극이 그와 동시에 발생하도록 조작하거나, 반대로 청각자극이 제시될 때 먼젓번 조건에서처럼 시각자극이 동시에 발생되도록 조작하는 두 조건에서 실험참여자가 운동제어를 통하여 자극 제시주기를 정확하게 예측하는지를 관찰하고자 하였다. 실험 결과, 청각자극에 반응하여 시각자극을 조작하는 조건에서 실험참여자들의 예측 반응 정확성이 높았으며, 자극 제시 주기를 500ms/800ms/2000ms 세 가지로 랜덤하게 제시했을 때 500ms에서 두 조건 모두 정확한 예측 반응을 보였으나 800ms 이상에서는 예측 반응의 정확성이 떨어져 반응 그래프가 양봉 분포를 보이는 것으로 확인되었다. 본 연구는 청각자극이 전달되는 속도가 시각자극이 전달되는 속도보다 빠르며 청각자극에 더 정확한

예측 반응을 보인다고 보고한 선행 연구의 결과들을 재확인하는 한편, 이와 같이 운동 제어가 매개하는 다중 감각 통합 상황에서의 예측 반응의 정확성 및 정밀성이 자극 제시 주기에 따라 변화한다는 점을 새롭게 주지한다.

주요어 : 다중감각통합, 시청각통합, 운동제어, 리듬인지, 시간적 예측
학 번 : 2017 - 25502