The Enlightenment of the Theory Model of Audio-Visual Integration to Human-Computer Interaction Technology

Hong Lu* and Chaochao Pan

1 School of Education, Lanzhou City University, and Research Center of Urban Social Psychology of Lanzhou City University, No. 11, Jiefang Road, Anning District, Lanzhou, China
2 School of Psychology, Northwest Normal University, No. 96, Anning Road, Anning District, Lanzhou, China
Email: 401515190@qq.com

Abstract. The visual and auditory systems are the most important information processing systems in the brain. We usually sense external stimuli in the way of audio-visual integration. In this process, “visual” and “auditory” nervous form two functional systems which are relatively independent and interactive. It is of great significance for the fusion of multi-sensor data and human-computer interaction technology to understand the theoretical principles of integrating visual and auditory cues and the neural mechanism of audio-visual integration. Here, we try to explore new methods of audio-visual integration and develop new potential of human-computer interface. Previous studies have shown that the principles of integrating visual and auditory cues mainly include spatio-temporal rule, inverse-effectiveness rule and reliability rule. On this basis, researchers have gradually developed a theoretical model of audio-visual integration—causal inference model. By clarifying the mechanism of audio-visual integration, it provides a theoretical basis for the multimodal interaction technology of human-computer interaction, so as to improve the naturalness and efficiency of human-computer interaction and overcome the shortcomings of traditional human-computer interaction.

1. Introduction

Human beings are endowed with a variety of senses, including vision, hearing, smell, etc. these sensory systems cooperate with each other and provide an effective way for us to understand the world [1]. Among them, the visual and auditory systems are the most important information processing systems in the brain. When people perceive and obtain the external environment, they usually get visual and auditory cues concurrently. The neural system unconsciously drives the integration of audio-visual information to better complete the information processing. In recent years, cognitive neuroscience studies also show that the superior temporal sulcus (STS) may be the main “high-order” audio-visual integration area. The auditory input through the auditory associated cortex and the visual stimulation that through the visual are integrated into STS, and then the STS modulates the activity of specific single sensory functional area through feedback projection [2].

Multi sensory perception is an important ability of human intelligence, and also an important reason why human intelligence is different from artificial intelligence. Recently, with the rapid development of artificial intelligence, machine vision, voice recognition, motion control and other fields have achieved major breakthroughs, and began to widely penetrate into the financial, customer service, medical, driverless and other industries. Exploring human-computer interaction and artificial...
intelligence technology from the perspective of audio-visual integration to maximize its potential and make it continuously approach human intelligence. Artificial intelligence refers to the use of machines to achieve all tasks that can only be achieved with the help of human intelligence. Its essence is based on the continuous progress of learning ability and reasoning ability, imitating the process of human thinking, cognition, decision-making and action. On the way to realize artificial intelligence, human-computer interaction is essential. Human-computer interaction refers to the process of information exchange between people and computers who use a certain dialogue language to complete certain tasks in a certain way, including information input by people to computers through input devices, information feedback provided by computers through calculation and output devices, etc. The traditional human-computer interaction is manual input. At present, the human-computer interaction means developed or researched include voice input, line of sight tracking input and compound multi-channel input. The multimodal technology of human-computer interaction is a focus in the competition of information industry, which is regarded as a key technology in the world. Whereas the key of multimodal human-computer interaction is the mechanism of multimodal information integration.

Based on the perspective of audio-visual integration, here, we explore its integration mechanism, so as to provide inspiration for the development of human-computer interaction technology, and further improve the ecology and efficiency of human-computer interaction.

2. The Theoretical Model of Audio-Visual Integration

Organisms mainly receive information through audio-visual channels, and the input of each sensory system will be integrated into the field of perception according to a certain theoretical model. In recent years, the empirical “rules” of integrated processing established in experimental research have laid a foundation for the quantitative model of audio-visual integration. Firstly, the “spatio-temporal rule” suggests that the stimulus presented concurrently or closely in time and space is more likely to be integrated [3]. Secondly, “inverse-effectiveness rule” predicts that when the information intensity of single sensory is lower, the integration effect is more significant [4]. Finally, according to the “reliability rule”, when integrating audio-visual signals, sensory with higher reliability (i.e., with lower signal to noise ratio (SNR)) are given higher weight [5]. On the basis of the above rules, researchers have gradually developed a theoretical model of audio-visual integration—Bayesian causal inference model.

According to causal inference model, human perception of the external world is based on probability inference based on Bayesian principle. From the Bayes point of perception, the brain combines the noise signal and prior knowledge to infer the environmental attributes best. In general, the formula $\hat{S}_i = f_i(S)$ is used to express the estimation of the neural system to the environmental attributes. Among them, $S$, $f_i$ subscript $i$ and $\hat{S}_i$ respectively represent the physical attributes to be estimated, the estimation operation conducted by the nervous system, the sensory and the final perceptual estimation. Under the condition of no prior information, assuming that the noises between senses are independent of each other and have Gaussian distribution, the error of the final estimation is the smallest. Specifically, the nervous system integrates sensory signals into a unified and coherent perception in the way of channel reliability weighting. In addition, the model also believes that human brain, as an ideal observer, can not only infer the position of signal source ($S$) from two sensory signals ($S_A$ and $S_V$), but also judge whether these signals have common cause (C). How the brain integrates information depends on the causal relationship it infers. However, due to the presence of sensory noise, the nervous system can not directly obtain the source location, but only the sensory signal noise measurements ($X_A$ and $X_V$) of the signal, so this inference becomes complicated. Considering a variety of possible causal structures, the model proposes two hypotheses: there is a common cause or there are different reasons between sensory cues. As shown the left side of figure 1, the brain infers that there may be a common cause ($C = 1$) for visual and auditory cues that co-exist in time and space. In this situation, the nervous system integrates and adds sensory cues according to the weight of different channel information. When there are significant differences between visual and
auditory cues, the brain infers that these cues can come from two different reasons (C = 2) (as shown on the right side of figure 1), so it processes them as two independent events or objects respectively. In each trial, the causal inference model considers the above two scenarios at the same time. The brain infers the probability of the causal structure with common cause (left, C = 1) and the probability of the causal structure with two independent causes (right, C = 2), and then weighted average them.

Figure 1. Schematic diagram of causal inference model [6].

Ventriloquism effect is the classic presentation of audio-visual integration. The reason for this phenomenon is that the subjects infer that there is a common cause (puppet speaking) between visual stimulation (puppet lip pronunciation action) and auditory stimulation (speech). On this basis, in order to test the explanatory power of the causal inference model, Körding et al. [7] used the improved ventriloquism effect as the experimental paradigm to present the participants with short visual (flash) and auditory stimulation (auditory Beep), but changed the spatial differences between them, requiring the subjects to report the positions of the perceived visual and auditory stimulation respectively by pressing keys. The experimental results are consistent with the predicted values of the model, and the causal inference model can well explain the experimental data ($R^2 = 0.97$).

In recent years, Kayser et al. [8] have combined behavior research with fMRI research of multivariate pattern decoding, trying to reveal how causal inference maps to specific brain regions. Results as shown in figure 2, the independent processing of sensory cues is reflected in the primary sensory cortex, the integrated processing is reflected in the posterior parietal region, and the anterior parietal region reflects the overall causal inference estimation. This distributed sensory representation model shows that causal reasoning is gradually carried out along the cortical hierarchy.

Figure 2. Causal inference map.

3. Enlightenment to Human-Computer Interaction Technology
Generally, things or events in environment always transmit a variety of information (such as voice, color, etc.) at the same time, and the nervous system can automatically integrate clues of different modes, and make their perception as a unified, continuous and stable whole. Multisensory integration can effectively make up for the lack of single-mode information and uncertainty, which is helpful for more rapid and accurate information acquisition. The same is also true for the setting of human-
computer interface. On the basis of eliminating redundancy effect, how to enrich human-computer interaction mode to improve its interactivity and efficiency has always been a problem in human-computer interface design. On the basis of summarizing the theoretical model of audio-visual integration, this paper aims to provide inspiration for the development of human-computer interaction technology.

First of all, the effective integration of “seeing” and “listening” is the starting point of multimodal human-computer interaction technology. With the current human-computer interaction involves more and more modalities, each modality not only affects each other, but also separates, the importance of audio-visual integration theory in the field of human-computer interaction is increasingly prominent.

As mentioned, the visual and auditory systems are the most important information processing systems in the brain. We usually sense external stimuli in the way of audio-visual integration. This enlightens us that in the human-computer interaction technology, we can combine the scan input of audio-visual integration to improve the sense of reality and immersion of human-computer interaction.

Secondly, it is the future development trend to extend the traditional human-computer interaction mode of key, knob, touch screen and other input interfaces to the integration of audio-visual dual-mode and even multi-mode human-computer interaction. Under the influence of scientific and technological progress and market demand, the traditional interaction mode of keys has gradually transformed into multi-modal human-computer interaction mode, such as touch control, voice, gesture recognition and so on. Multimodal interaction is more efficient, more in line with the human body’s natural habits and perception needs, and the task completion rate is higher. On the basis of clarifying the mechanism of audio-visual integration and multi-sensory integration, it is applied to human-computer interaction technology.

Finally, speech interaction based on visual presentation is the future direction. At present, speech interaction based on visual presentation is one of the important research directions. The content displayed by the machine is through the screen (vision), and the input is mainly voice interaction. This way can provide users with an efficient, convenient and comfortable human-computer interaction system. On the basis of scientific research and technological progress, when human-computer interaction is as natural as human communication, the real era of intelligence will come.

4. Conclusion
Based on the perspective of audio-visual integration, we summarized the theoretical models and corresponding neural mechanisms of popular audio-visual integration, in order to provide a theoretical basis for the development of human-computer interaction technology.

Acknowledgments
This work was supported by the Ethnic Research Project of the State People’s Committee Foundation of China (2019-GMD-020).

References
[1] Zhou A B, Pan C C, Jin G and Lu H 2020 Enlightenment of audiovisual integration theory model on hearing training for hearing-impaired people Chinese Journal of Special Education 2 21-8.
[2] Zhou A B and Pan C C 2020 The significance of audiovisual integrated neurocognitive mechanism in the study of multimodal perception of hearing-impaired people Journal of Audiology and Speech Pathology 28 1-6.
[3] Chandrasekaran C 2017 Computational principles and models of multisensory integration Current Opinion in Neurobiology 43 25-34
[4] Colonius H and Diederich A 2018 Formal models and quantitative measures of multisensory integration: a selective overview European Journal of Neuroscience 8 1-44.
[5] Ohshiro T, Angelaki D E and DeAngelis G C 2011 A normalization model of multisensory integration Nature Neuroscience 14(6) 775-782.
[6] Andersen T S 2012 Predictive power in models of audiovisual integration of speech Seeing and Perceiving 25 105.

[7] Körding K P and Wolpert D M 2006 Bayesian decision theory in sensorimotor control Trends in Cognitive Sciences 10 319-26.

[8] Kayser C and Shams L 2015 Multisensory causal inference in the brain Plos Biology 13 1-15.