What Can Computational Models Learn From Human Selective Attention? A Review From an Audiovisual Unimodal and Crossmodal Perspective

Di Fu, Cornelius Weber, Guochun Yang, Matthias Kerzel, Weizhi Nan, Pablo Barros, Haiyan Wu, Xun Liu and Stefan Wermter

1. INTRODUCTION

"The art of being wise is knowing what to overlook."

–William James, 1842-1910.

The real world is complex, uncertain and rich in dynamic ambiguous stimuli. Detecting sudden changes in the environment is significant for organisms to survive because these events need prompt identification and response (Todd and Van Gelder, 1979). Considering the limited capacity for processing information, selective attention is like a filter with the ability to remove unwanted or irrelevant information and thus optimizes a human's action to achieve the current goal (Desimone and Duncan, 1995). It is crucial as well for intelligent agents to integrate and utilize external and
internal information efficiently and to reach a signal-to-noise ratio as high as humans can (signal detection theory, SDT) (Green and Swets, 1966; Swets, 2014).

Selective attention is involved in the majority of mental activities, and it is used to control our awareness of the internal mind and the outside world. Selective attention also helps to integrate information from multidimensional and multimodal inputs (Talsma et al., 2010). Empirical research shows that stimuli with multimodal properties are more salient than unimodal stimuli; therefore, selective attention is more easily captured by multimodal inputs to promote further processing (Van der Burg et al., 2008, 2009). Selective attention is predominantly categorized by psychologists and neuroscientists into “endogenous” and “exogenous” attention. Endogenous attention helps to allocate limited cognitive resources to the current task (Posner and Snyder, 1975; Corbetta and Shulman, 2002; Styles, 2006). The metaphor for this process is described as directing a spotlight in a dark room. Such a process helps us, for instance, to search for one specific email only by glimpsing the crammed email box. However, the action can sometimes be interrupted by attractive advertisements or breaking news on a website. This latter kind of orienting attention is called exogenous attention which is usually caused by an unexpected change in the environment. It is considered to be instinctive and spontaneous and often results in a reflexive saccade (Smith et al., 2004; Styles, 2006). Another point of view distinguishes between “covert” and “overt” orienting attention: covert attention can attend events or objects with the absence of eye movement, while overt attention guides the fovea to the stimulus directly with eyes or head movements (Posner, 1980). This is because covert attention requires inhibition of saccades to sustain fixation, which is not needed during overt attention (Kulke et al., 2016). Analogously, covert and overt mechanisms exist in the auditory system. Since humans cannot move ears like eyes, the difference between these two mechanisms is that covert auditory attention can govern attention without any motion, while overt auditory attention attends to sound sources with head movements (Kondo et al., 2012; Morillon and Baillet, 2017). Head movements contribute to sound localization during overt auditory attention (Wallach, 1940; Perrett and Noble, 1997).

To understand the mechanisms underlying selective attention is helpful for computational models of selective attention for different purposes and requirements (Das et al., 2017). Attention models have been proposed and applied in computer science for decades, and attention mechanisms have achieved high performance in sequence modeling (Vaswani et al., 2017; Peng et al., 2019). Bio-inspired implementations of attention in computer science address the limited computation capacity of machines through assigning computational resources by priority (Xu et al., 2015). However, gaps exist between computational models and theories of human selective attention. Some theories are metaphysical and mystifying, especially for readers that lack experience in humans’ behavioral and neural studies. Frintrop et al. (2010) published a survey about computational visual systems with an extensive description of the concepts, theories and neural pathways of visual attention mechanisms. It is stated that “the interdisciplinarity of the topic holds not only benefits but also difficulties: concepts of other fields are usually hard to access due to differences in vocabulary and lack of knowledge of the relevant literature” (p. 1). These interdisciplinarity challenges are still unsolved thus far. Additionally, the development and application of technical measurements and methods like functional magnetic resonance imaging (fMRI), Magnetoencephalography (MEG), and state-of-the-art artificial neural networks (ANN) and deep learning (DL) open up a new window for studies on humans, primates, and robots. Such new findings should be valued and integrated into the current framework.

Although there are several review articles on selective attention in the field of both psychology and computer science (Shinn-Cunningham, 2008; Frintrop et al., 2010; Lee and Choo, 2013), most of them only focus either on a single modality or on general crossmodal processing (Lahat et al., 2015; Ramachandram and Taylor, 2017). However, it is essential to combine and compare selective attention mechanisms from different modalities together to provide an integrated framework with similarities and differences among various modalities. In the current review, firstly, we aim to integrate selective attention concepts, theories, behavioral, and neural mechanisms studied by the unimodal and crossmodal experiment designs. Secondly, we aim to deepen the understanding of the interdisciplinary work in multisensory integration and crossmodal learning mechanisms in psychology and computer science. Thirdly, we aim to bridge the gap between humans’ behavioral and neural patterns and intelligent system simulation to provide theoretical and practical benefits to both fields.

The current review is organized into the following parts. Section 2 is about the existing mainstream attention theories and models based on human experimental findings and attention mechanisms in computer science. Section 3 summarizes human visual selective attention studies and introduces the modeling work in computer science inspired by psychology. Section 4 describes results on less studied auditory selective attention and the corresponding modeling work. Section 5 reviews mechanisms and models about crossmodal selective attention and state-of-the-art approaches in intelligent systems. Here, to provide focus, we select the most representative phenomena and effects in psychology: Pop-out Effect (visual attention), Cocktail Party Effect (auditory attention), and audiovisual crossmodal integration and conflict resolution (crossmodal attention). Since these effects are also well-established and often simulated in computer science, we highlight the classic and latest work. Finally, we discuss the current limitations and the future trends of utilization and implications of human selective attention models in artificial intelligence.

2. DIFFERENT THEORIES AND MODELS OF SELECTIVE ATTENTION

2.1. Classic Bottom-Up and Top-Down Control vs. Priority Map Theory

The mainstream view of selective attention proposes that there exist two complementary pathways in the brain cortex, the
dorsal and ventral systems. The former, which includes parts of the intraparietal sulcus (IPS) and frontal eye field (FEF), is in charge of the top-down process guided by goals or expectations. The latter, which involves the ventral frontal cortex (VFC) and right temporoparietal junction (TPJ), is in charge of the bottom-up process triggered by sensory inputs or salient stimuli. Other research finds that emotional information can also capture attention. For instance, selection history means the potential stimuli of interest, and anatomical mechanisms of attention play a dominant role in planning, decision-making, conflict detection and resolution. The anterior cingulate cortex (ACC) and lateral prefrontal cortical regions are involved in the executive control component (Benes, 2000) and the orienting function is responsible for selecting task-related or survival-related information from all the sensory inputs. The orienting network also determines an attention shift between exogenous attention engagement (bottom-up) and endogenous attention disengagement (top-down).

### 2.2. Functional Neural Networks Model

The Functional neural networks model separates attention into clear sub-components. Fan and Posner designed the Attentional Network Test (ANT) by combining the classic Flanker task and Posner cueing task to provide a quantitative measurement for studying the sub-components: alerting, orienting, and executive control (Fan et al., 2002, 2005; Fan and Posner, 2004). The component of the alerting network increases the focus on the potential stimuli of interest, and anatomical mechanisms of alerting are correlated with the thalamic, frontal, and parietal regions. The orienting function is responsible for selecting task-related or survival-related information from all the sensory inputs. The orienting network also determines an attention shift between exogenous attention engagement (bottom-up) and endogenous attention disengagement (top-down). Alerting is associated with the superior parietal lobe (SPL), TPJ, and frontal eye fields (FEF). The executive control component of attention plays a dominant role in planning, decision-making, conflict detection and resolution. The anterior cingulate cortex (ACC) and lateral prefrontal cortical regions are involved in the executive control component (Benes, 2000). During the ANT, participants are asked to determine the direction of the central arrow above or below the fixation. The central arrow is accompanied by congruent or incongruent flankers. In neutral conditions, the central arrow has no flankers. There are four cue conditions: no cue, center cue, double cue, and spatial cue. Effects are calculated by subtracting participants’ reaction time (RT) under two different conditions: the alerting effect = RT (no-cue) - RT (double-cue); the orienting effect = RT (center cue) - RT (spatial cue); the executive control effect = RT (incongruent flanking) - RT (congruent flanking) (Fan et al., 2002). Clinical studies using the ANT can explore specific differences of cognitive performance between patients and healthy participants (Urbanek et al., 2010; Togo et al., 2015). For example, Johnson et al. (2008) used the ANT to test children with attention deficit hyperactivity disorder (ADHD) and found that they show deficits in the alerting and executive control networks but not in the...
orienting network. The model and findings arising from the ANT could serve to provide useful interventions for clinical treatment.

2.3. Neural Oscillation Model

Neural oscillations characterize the electrical activity of a population of neurons (Musall et al., 2012). Synchronization of oscillations is the coordination of firing patterns of groups of neurons from different brain areas (Varela et al., 2001). In contrast, the desynchronization of oscillations is the inhibition of neuron activities with opposite phases. Attention is correlated with synchronization and desynchronization of specific cortical neural oscillations. Clayton et al. (2015) propose a gamma-theta power-phase coupling model of attention and point out that attention is selectively adjusted via the excitation of task-relevant processes and the inhibition of task-irrelevant processes (see Figure 1B). The excitation of task-relevant processes is controlled by frontomedial theta (fm-theta) power (4–8 Hz) from the posterior medial frontal cortex (pMFC) to the lateral prefrontal cortex (LPFC). Among the communication between LPFC and excited sensory areas, gamma power (>30 Hz) is associated with the excitation of the task-relevant processes. The inhibition of task-irrelevant processes is linked with alpha power (8–14 Hz). The pMFC deploys the crucial inhibition processing by controlling the default mode network [posterior cingulate cortex (PCC) and ventromedial prefrontal cortex (vmPFC)] via the alpha oscillation. The limitation of the model is that the results obtained and presented across different brain regions are mainly correlations and descriptive results rather than causal relationships. Besides, most of the empirical evidence for the model was obtained by visual tasks instead of other modalities. Nevertheless, this gamma-theta power-phase coupling model shows interpretative neural pathways of the neural oscillation of selective attention.

2.4. Free-Energy Model and Information Theory

The free-energy model explains attention from a hierarchical inference perspective (Friston, 2009; Feldman and Friston, 2010). The gist of the model is that the stimuli in the living environment can be viewed as sensory inputs, surprise or uncertainty which can increase the entropy of the human brain. Our brains have a tendency to maintain the information order to minimize the energy cost caused by surprise. In doing so, perception brings about the sensory inputs, and attention infers the consequence caused by the inputs to adjust action and control the entropy growth.

Corresponding to the free-energy model, Fan's review (Fan, 2014) tries to combine the information theory and experimental neural findings to explain the top-down mechanisms of humans' cognition control (the hub of the cognition capacity) and selective attention. Inspired by the free-energy view, Fan points out that cognitive control is a high-level uncertainty or entropy reduction mechanism instead of a low-level automatic information perception. According to Shannon's information theory (Shannon, 1948), uncertainty can be quantified by entropy, and the rate of entropy is used to calculate the time density of the information transmission through different channels. Performance costs appear during cognitive channel switching. The benefits of the information theory are that attention or other cognitive processes can be quantified, and situations (like incongruent or congruent conditions in conflict processing) can be computed as bits quantitatively. Fan assimilates stimulus types, time frequency of the stimulus presentation, and human reaction time from cognitive psychology experimental tasks into entropy, surprise, and channel capacity. In this theory, if we know the probability of an event or a stimulus condition, we can calculate the surprise value of that condition and infer the information processing rate. For example, studies found that visual attention can select 30–60 bits per glimpse (Verghese and Pelli, 1992) and the upper limit of human information processing is around 50 bps. Under this framework, the anterior insula (AI) and the anterior cingulate cortex (ACC) are associated with processing the uncertain inputs and the frontoparietal cortex plays a ubiquitous role in the active control.

Research from network neuroscience takes a similar viewpoint that the brain is designed to be functioning with the lowest cost (Bullmore and Sporns, 2012; Barbey, 2018). However, the free-energy model and information theory concentrate on top-down control pathways which may fail to explain some bottom-up phenomena. For instance, why can human attention be captured by the salient external stimuli involuntarily? It can cause the rise of the information entropy and be opposite to the hypothesis that the human brain instinctively resists the disorder. Besides, experimental evidence of processing channels is still lacking.

2.5. Attention Mechanisms in Computer Science

Previous models (1980s–2014) mainly use the saliency-based winner-take-all algorithm based on human datasets to mimic humanlike visual or auditory attention (Borji and Itti, 2012; Lee and Choo, 2013). Those models aim to extract the target information from the environment or noisy background. In recent years since 2014, attention mechanisms have been applied to Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long-short Term Memory (LSTM) for sequence modeling work. Attention mechanisms were firstly used in computer vision (Ba et al., 2014) and then became widely used across different domains according to the type of input data, such as object recognition (Hara et al., 2017), image description generation (Xu et al., 2015), speech recognition (Chorowski et al., 2015), machine translation (Luong et al., 2015), video caption generation (Gao L. et al., 2017), sentiment classification (Wang et al., 2016), visual question answering (Li et al., 2018), etc.

Attention mechanisms in computer science can be distinguished as soft and hard attention (Xu et al., 2015), or as global and local attention (Luong et al., 2015). Soft attention is the expectation of selected information in the input attention distribution. For example, there is a translation task to translate one German sentence “Ich komme aus Deutschland” into an English sentence “I come from Germany.” In machine translation, attention scores mean different weights assigned to words in the source sentence (German) according to each word
TABLE 1 | Main theories of visual selective attention based on various processing pathways.

| Theory                                      | Viewpoint                                                                 | Processing              |
|---------------------------------------------|---------------------------------------------------------------------------|-------------------------|
| Stimulus-driven Theory (1992)               | Singletons automatically capture visual attention                        | Bottom-up               |
| Goal-driven Theory (1992)                   | Individuals’ intentions determine attentional capture                     | Top-down                |
| Contingent Capture Hypothesis (1992)        | Contingent on attentional control settings induced by task demands        | Top-down                |
| Attention Selection Bias Competition (1995) | Response to distractors around the target is inhibited                    | Bottom-up & Top-down    |
| Signal Suppression Hypothesis (2010)        | Salience signal automatically generated by singletons can be suppressed   | Bottom-up & Top-down    |

in the target sentence (English). In this example, corresponding to “Germany,” “Deutschland” should be assigned more weights than other words in the source sentence. Soft attention focuses more broadly than hard attention. Hard attention only concentrates on information of the specific location by assigning zero weight to other information (Xu et al., 2015). The concepts of global and local attention vaguely correspond to soft and hard attention, respectively. Recently, an important application is the self-attention mechanism (Vaswani et al., 2017). Different from soft and hard attention, self-attention does not capture features of mapping between source and target but can learn the inherent structure both within the source and target text. In the above example, “from” is more likely to be followed by “Germany.” Self-attention can be applied in each decoder layer of neural networks to achieve distributed processing (Bahdanau et al., 2014). In this way, self-attention shows good performance and efficiency when the input sentence is too long as in machine translation (Luong et al., 2015) or the input image is too large as in computer vision (Peng et al., 2019).

In summary, we conclude in this section that human attention is a process to allocate cognitive resources with different weights according to the priority of the events. Similarly, in computer science, attention mechanisms in models are designed to be allocating different weights to relevant input information and ignore irrelevant information with low-valued weights. However, the connection between computer science models and psychology is still loose and broad. Especially for understanding crossmodal selective attention from a functional view, it is required to explore the human cognition processing from a computational perspective, which is also beneficial for confirming psychological and biological hypotheses in computer science.

3. VISUAL SELECTIVE ATTENTION—“POP-OUT” EFFECT

3.1. Behavioral and Neural Mechanisms of Human Visual Selective Attention

Many systematic reviews in the areas of primate vision and computer vision have introduced the concepts and research findings in visual selective attention (Frintrop et al., 2010; Borji and Itti, 2012; Lee and Choo, 2013). In our current review, we further concentrate in particular on mechanisms of the “pop-out” effect and computational models based on the saliency map. In general, the “pop-out” effect describes saliency processing. Considering that an object is not salient by itself (Itti and Baldi, 2009), the “pop-out” effect usually happens when an object has more salient physical features than other objects in the context, such as location, color, shape, orientation, brightness, etc. (VanRullen, 2003). Saliency can also be extended to affective and social domains, like familiarity, threat, etc. (Fan, 2014). Humans’ attention can be immediately captured by salient objects, which can explain why the warning signs on streets are always red and apparent.

Nevertheless, controversy remains about the role of top-down control when a salient stimulus captures attention. Stimulus-driven theory (bottom-up saliency hypothesis) suggests that an abrupt-onset object can automatically capture humans’ attention without any intention and be processed faster than other non-onset elements (Yantis and Jonides, 1984; Theeuwes, 1991). To the contrary, the goal-driven theory (Bacon and Egeth, 1994) and the contingent capture hypothesis (Folk et al., 1992) propose that the overlap dimension between stimulus properties and task setting goals is the crucial factor, since it can determine whether the salient stimulus can be captured or not. Experiments show that if the salient stimulus has no task-relevant feature, participants adopt a feature-search mode autonomously to suppress the distraction from the salient stimulus (Bacon and Egeth, 1994).

Hybrid theories attempt to integrate components of both stimulus-driven and goal-driven theories in attention capture. Findings from monkey studies showed that attention selection through biased competition occurred when the target and the distractor were both within the receptive field. Neurons responded primarily to the target, whereas the responses to the distractor were attenuated (Desimone and Duncan, 1995). Subsequently, Mounts (2000) discovered a phenomenon named “surround inhibition.” If a salient stimulus appears near the target, it can be inhibited by top-down control. Later, the signal suppression hypothesis proposed that the salient stimulus automatically generates a salience signal at first and then the signal can be subsequently suppressed, possibly resulting in no attention capture (Sawaki and Luck, 2010; Gasperlin et al., 2015, 2017) (the theories are summarized in Table 1).

Neural findings of humans and primates contribute a lot to understand saliency processing in the primary cortex and subcortex. The saliency map theory (Li, 1999, 2002) suggests that neurons in the primary visual cortex (V1) play a crucial role for the input feature processing during the “pop-out” effect. V1 is the neural foundation of the preattentive process during visual search, and it only responds to stimuli located in the
classical receptive fields (CRFs). In this saliency map theory, V1 is considered to define the saliency degree of visual inputs. Various features of the target and context enter into the V1 CRFs at the same time. When features of the target are more significant than the context, the target pops out. The saliency map computes the saliency value for all locations in the CRFs rather than only encoding the target location (Veale et al., 2017). In comparison to the classical feature integration model (Treisman and Gormican, 1988) and Itti’s saliency model (Itti and Koch, 2000), the main property of the saliency map theory is that saliency processing is only based on a single general feature selection map rather than using a combination map to bind several individual feature maps together. Furthermore, dominant inputs from V1 convey signals to an evolutionarily old structure in the midbrain—the superior colliculus (SC). Superficial layers of the SC encode saliency representations through center-surround inhibition and transfer the inputs to deep layers to trigger priority selection mechanisms to guide attention and gaze (Stein et al., 2002; Veale et al., 2017; White et al., 2017). There is not only bottom-up processing in the primary visual cortex and SC, but also top-down processing. Within the primary visual cortex, the top-down mechanism is mediated by V2 and the interaction occurs in human V4 (Melloni et al., 2012). Moreover, deep layers of the SC represent goal-related behaviors independent of the visual stimuli (Hafed and Krauzlis, 2008; Hafed et al., 2008; Veale et al., 2017).

The large-scale human brain networks also play important roles in visual selective attention. The salience network (SN), composed of AI (anterior insula) and ACC (anterior cingulate cortex), is considered to be working as the salience filter to accept inputs from the sensory cortex and trigger cognitive control signals to the default mode network (DMN) and central-executive network (CEN). Functions of the SN are mainly about accomplishing the dynamic switch between externally and internally oriented attention (Uddin and Menon, 2009; Menon and Uddin, 2010; Uddin, 2015). Another taxonomic cingulo-opercular network shares a large overlap with the SN, containing the anterior insular/operculum, dorsal anterior cingulate cortex (dACC), and thalamus. The cingulo-opercular network has the highest cortical nicotinic acetylcholine receptor (nAChR) density, which is highly correlated with attention functions (Picard et al., 2013). However, conclusions about functions of the cingulo-opercular network are not consistent. For instance, Sadaghiani and D’Esposito (2014) revealed that the cingulo-opercular network plays a role in staying alert but not in selective attention during visual processing. In sum, the V1 and SC consist of primary cortex-subcortex pathways of saliency processing and attention orienting. The AI and ACC consist of large-scale functional networks of saliency processing, alertness and attention shifting. However, the correlation or interaction between these two pathways remains unclear.

Besides elementary physical salient features, scene regions that contain semantic meaning also proved to play a critical role in attentional guidance (Henderson and Hollingworth, 1999; Wolfe and Horowitz, 2017). Henderson and Hayes (2017) express the spatial distribution of meaning across scenes as meaning maps, which are obtained by participants’ ratings of the meaningfulness of scene regions. They encode the meaning maps comparable to the image salience and operationalize the attention distribution to be duration-weighted fixation density. Their work demonstrates that both, salience and meaning, predict attention but only meaning guides attention while viewing real-world scenes. According to the cognitive-relevance theory of attentional guidance, the meaning maps contain more semantic information for the real context. Their updated findings appear to be particularly insightful and practical for artificial intelligence methods for labeling real-world images.

### 3.2. Computational Models Based on Human Visual Selective Attention

Based on human saccade and fixation research, a vast body of bio-inspired visual attention models has been developed and broadly applied in object segmentation (Gao G. et al., 2017), object recognition (Klein and Frintrop, 2011), image caption generation (Bai and An, 2018), and visual question answering (VQA) (Liu and Milanova, 2018). The visual attention model aims to predict the human eye fixation with minimal errors (Borji and Itti, 2012). Consistent with humans’ visual processing pathways, models in visual attention are generally classified based on the bottom-up and top-down streams (Borji and Itti, 2012; Liu and Milanova, 2018). Bottom-up models are successful in modeling low-level and early processing stages (Khaligh-Razavi et al., 2017). The most classic saliency model, which uses features of color, orientation, edge, and intensity, allocates an attention weight to each pixel of the image (Itti et al., 1998; Itti and Koch, 2000) (see Figure 2A). The “winner-take-all” strategy is the core algorithm of saliency models. However, several criticisms on the saliency model cannot be ignored either. For instance, a salient feature is obtained by calculating the difference between input at one location and other input surrounding it so that any spatial discontinuities of features can be detected (Itti et al., 1998). This center-surround scheme is analogous to attention selection via bias competition within the visual receptive fields (Desimone and Duncan, 1995). However, the salient feature obtained by this scheme can only correspond to a small local region of an image scene with higher contrast but not to a whole object or an extended part of it (VanRullen, 2003; Lee and Choo, 2013) (also see Figure 2A).

In contrast, high-level task-driven attention models remain to be explored and developed further. Some research predicts human eye fixation with free-viewing scenes based on end-to-end deep learning architectures (Jetley et al., 2016; Kruthiventi et al., 2017; Kummerer et al., 2017). Deep neural networks (DNNs) have sometimes been shown to have better performance than other known models by using top-down processing mechanisms. Especially, DNNs can successfully simulate human-like attention mechanisms (Hanson et al., 2018). Here task-driven components can not only be implemented as targets but also implemented as prior knowledge, motivation, and other types of cues. Furthermore, models like DeepFeat incorporating bottom-up and top-down saliency maps by combining low- and high-level visual factors surpass other individual bottom-up and top-down approaches (Mahdi et al., 2019). Nowadays, computer vision research intends to make models learn the semantic
meaning rather than simply classify objects. For instance, image captioning requires models not only to detect objects but also extract relationships between objects (Hinz et al., 2019). Co-saliency tends to be a promising preprocessing step for many high-level visual tasks such as video foreground extraction, image retrieval, and object detection. Because co-saliency implies priorities based on human visual attention, it can detect the most important information among a set of images with a reduced computational demand (Yao et al., 2017). In future research, co-saliency approaches may be combined with the meaning maps of human attention for better image interpretation accuracy.

As the number of interdisciplinary studies keeps increasing, research from psychology and artificial intelligence complement each other deepening the understanding of human visual attention mechanisms and improving the performance of computational models. On the one hand, psychologists interpret humans’ behavioral or neural patterns by comparing them with the performance of DNNs. For example, Eckstein et al. (2017) found that human participants often miss giant targets in scenes during visual search but computational models such as Faster R-CNN (Ren et al., 2015), R-FCN (Dai et al., 2016), and YOLO (Redmon and Farhadi, 2017) do not show any similar recognizing failures. Their results suggest that humans use “missing giant targets” as the response strategy to suppress potential distractors immediately. On the other hand, computer scientists interpret features of computational models by comparing their performance with simulations of humans’ behaviors. For instance, Hanson et al. (2018) found that the Deep Learning (DL) network rather than the single hidden layer backpropagation neural network can replicate human category learning. This is because hidden layers of the DL network can selectively attend to relevant category features as humans do during category learning.

4. AUDITORY SELECTIVE ATTENTION—COCKTAIL PARTY EFFECT

4.1. Behavioral and Neural Mechanisms of Human Auditory Selective Attention

At a noisy party, a person can concentrate on the target conversation (a top-down process) and easily respond to someone calling his/her name (a bottom-up process). This capability (in a real-life scenario) is named “Cocktail Party Effect” (Cherry, 1953). Auditory information conveys both temporal and spatial features of objects. For instance, we can determine whether water in a kettle is boiling by the special sounds of different heating phases. Auditory scene analysis (ASA) allows the auditory system to perceive and organize sound information from the environment (Bregman, 1994). Since humans cannot close their ears spontaneously to avoid irrelevant information, selective attention is important to segregate the forefront auditory information from a complex background and distinguish meaningful information from noise. Besides, auditory selective attention allows humans to localize sound sources and filter out irrelevant sound information effectively.

In the Cocktail Party problem, energetic masking and informational masking cause ambiguity between the auditory target and noise in the environment. Energetic masking occurs when different sound sources have overlaps in frequency spectra at the same time. The perception and recognition of the target sound can be weakened physically by noise (e.g., the target speech

![Figure 2](image-url)

**Figure 2** | (A) Visual saliency model. Features are extracted from the input image. The center-surround mechanism and normalization are used to generate the individual feature saliency maps. Finally, the saliency map is generated by a linear combination of different individual saliency maps (adapted from Itti et al., 1998); (B) Auditory saliency model. The structure of the model is similar to the visual saliency model by converting sound inputs into a frequency “intensity image” (adapted from Kayser et al., 2005).
overlaps with a white noise masker). Informational masking occurs when the target and masker voices sound similar (e.g., a target male is speaking while another non-target male is speaking at the same time). The listener cannot discriminate them perceptually (Brungart, 2001; Lidestam et al., 2014). The neural mechanisms of these two causes are different. Scott et al. (2004) asked participants to listen to a target speaker with added noise (energetic masking) or added speech (informational masking). They found that informational masking was associated with the activation in the bilateral superior temporal gyri (STG) and energetic masking was associated with the activation in the frontoparietal cortex. The activation was correlated with explicit attentional mechanisms but not specifically to the auditory processing.

In accordance with the Gestalt framework, ASA is the solution to the Cocktail Party problem (Bee and Michely, 2008). Similar to visual processing, ASA can be separated into two components. The primitive analysis (bottom-up process) and the schema-based processing (top-down process) (Bregman, 1994). In the primitive analysis, auditory signals are separated into independent units and integrated into disparate auditory streams according to sound features and time-frequency. In the schema-based processing, prior knowledge such as language, music, other auditory memory, and endogenous attention helps to compare the auditory input signals with previous experience (Shinn-Cunningham, 2008) (see Figure 3A). In laboratory studies, psychologists adopt the dichotic listening paradigm to mimic the Cocktail Party problem. During the task, participants are asked to attend to the auditory materials presented to one ear and ignore the auditory materials presented to the other ear. Afterwards, participants are asked to report the information from the attended or unattended ear. Previous studies show that a higher working memory capacity (WMC) predicts a better attention focus (Conway et al., 2001; Coliflesh and Conway, 2007), because a lower capacity cannot accomplish segregation and grouping of any auditory information well. Those findings are in accordance with the controlled attention theory of working memory (Baddeley et al., 1974).

Event-related potential (ERP) N1-P2 components, alpha oscillations, and frequency-following responses (FFRs) disclose how the human brain copes with the Cocktail Party problem (Du et al., 2011; Strauß et al., 2014; Lewald and Getzmann, 2015). The ERP N1 component peaks between 80 and 120 ms after the onset of a stimulus. It is sensitive to the exogenous auditory stimuli features (Michie et al., 1990). N1 (equivalent in MEG is M100) is generated from the primary auditory cortex (A1) around the superior surface of the temporal lobes (Zouridakis et al., 1998). P2 is always observed as the following component of N1. It peaks at around 200 ms after receiving the external stimulus. These early components support the early selection model of auditory attention (Woldorff et al., 1993; Broadbent, 2013; Lee et al., 2014). Alpha oscillations occur in the parietal cortex and other auditory cortical regions during spatial attention. Selective attention modulates alpha power oscillations in temporal synchrony with the sensory input and enhances the neural activity related to attended stimuli. Wöstmann et al. (2016) conducted a MEG study with a dichotic task and revealed that alpha oscillations are synchronized with speech rates and can predict the listener's speech comprehension. Scalp-recorded frequency-following responses (FFRs) are part of auditory brainstem responses (ABR). They are evoked potentials generated from the brainstem area (Mai et al., 2019). FFRs are phase-locked to the envelope or waveform of the low-frequency periodic auditory stimuli (Zhang and Gong, 2019). In the Cocktail Party problem, FFRs encode important features of speech stimuli to enhance the ability to discriminate the target stimuli from the distracting stimuli (Du et al., 2011). In summary, to exert the auditory selective attention, N1-P2 components are involved in perceiving and detecting the auditory stimuli in the early control processing; alpha oscillations and FFRs are mainly modulated by the selective control to accentuate the target and suppressing noise.

Analogous to the specialized streams of visual selective attention, there are “what” and “where” pathways in the auditory cortex (see Figure 3B). The ventral “what” pathway, which involves the anterolateral Heschl' gyrus, anterior superior temporal gyrus, and posterior planum temporale, is in charge of identifying auditory objects. The dorsal “where” pathway, which involves the planum temporale and posterior superior temporal gyrus (pSTG), is in charge of spatially localizing auditory objects. Within the “what” pathway, the supratemporal plane-inferior parietal lobule (STP-IPL) network dynamically modulates auditory selective attention; within the “where” pathway, the medial pSTG shows a higher-level representation of auditory localization by integrating the sound-level and timing features of auditory stimuli (Higgins et al., 2017; Häkkinen and Rinne, 2018). In addition, the “where” pathway is observed to activate around 30ms earlier than the “what” pathway implying that top-down spatial information may modulate the auditory object perception (Alain et al., 2001; Ahveninen et al., 2006). However, current studies find that functional overlaps exist in brain areas under different processing pathways, suggesting that brain areas are not function-specific (Schadwinkel and Gutschalk, 2010; Yin et al., 2014). The observed brain activities are not only stimulus-dependent but also task-dependent (Häkkinen et al., 2015). Besides, a suggested “when” pathway for temporal perception (Lu et al., 2017) deserves to be studied further because the temporal coherence is crucial for binding and segregating features into speech and speaker recognition when attention is engaged. Apart from the parallel pathways, the distributed processing under different structures may also provide feedback to facilitate the auditory attention (Bizley and Cohen, 2013).

For the Cocktail Party problem, previous neural findings show the attentional selective mechanism occurs in different phases of information processing. Ding and Simon (2012) found that the selective mechanism exists in both top-down modulation and bottom-up adaptation during the Cocktail Party problem. When the unattended speech signals were physically stronger, attended speech could still dominate the posterior auditory cortex responses by the top-down execution. Besides, when the intensity of the target was more than 8dB louder than the background, the bottom-up neural responses only adjusted to the target speaker rather than the background.
speaker. Golumbic et al. (2013) demonstrate that the selective mechanism happens only in the high-level cortices such as the inferior frontal cortex, anterior and inferior temporal cortex, and IPL. Here, only attended speech was selectively retained. However, in the low-level auditory cortices like the STG, both attended and unattended speech were represented. In addition, one study used functional near-infrared spectroscopy (fNIRS)-hyperscanning and found that the brain-to-brain interpersonal neural synchronization (INS) selectively enhances at the left TPJ only between the listener and the attended speaker but not between the listener and the unattended speaker. The listener's brain activity overtakes the speaker's showing a faster speech prediction by the listener. Besides, the INS increased only for the noisy naturalistic conversations with competing speech but not for the two-person conversation and was only associated with the speech content. Their findings implied that the prediction of the speaker's speech content might play an important role in the Cocktail Party Effect (Dai et al., 2018).

In summary, the human brain's auditory processing during the Cocktail Party problem is not hierarchical but heterarchical, which is mainly a bottom-up process aided by top-down modulation (Bregman, 1994). This includes interactions between different pathways and adaptations to the environment (Shinn-Cunningham, 2008; Bizley and Cohen, 2013).

4.2. Computational Models for the Human Cocktail Party Problem Solution

In the future, we may have moving robots offering food and drinks in noisy restaurants by precisely localizing speaking customers. Steps to solve the Cocktail Party problem in computer science can be mainly separated into: speech separation, sound localization, speaker identification, and speech recognition. The aims of acoustic models for the Cocktail Party problem are: identifying multiple speakers and disentangling each speech stream from noisy background. Numerous classical acoustic models are data-driven and based on algorithms of signal processing (Dávila-Chacón et al., 2018). Those models are robust and with good accuracy but lack the prior knowledge, biological plausibility and rely on the large datasets. Currently, models inspired by the human auditory attention system rely on smaller datasets and have shown improved adaptation. In this section, we focus on the following bio-inspired models: (1) computational auditory scene analysis (CASA): neural oscillator models as examples; (2) saliency models; (3) top-down- and bottom-up-based models.

Based on the Gestalt framework (Rock and Palmer, 1990), the goal of most CASA models is to segregate sounds with similar patterns or connections and group them into independent streams from the mixed auditory scene. Stemming from CASA models, neural oscillator models show good adaptation in auditory segregation. Neural oscillator models perform stream segregation based on the oscillatory correlation. Attention interest is modeled as a Gaussian distribution across the attended frequency. The attentional leaky integrator (ALI) consists of the connection weights between oscillators and the attentional process. The synchronized oscillators activate the ALI to separate sounds into streams like the endogenous attention mechanism (Wrigley and Brown, 2004). Furthermore, to make use of the temporal proximity of sound sources, Wang and Chang (2008) propose a two-dimensional (time and frequency) network
Wang and Terman, 1995 propose a model by extracting frequency, amplitude, and position as features and connecting them with the oscillator model LEGION. Segments with consistent features are bound into the saliency map according to the temporal correlation. Notably, a module of inhibition-of-return (IOR) is inserted to inhibit attention from fixing at the most salient scene for a long time. This mechanism can achieve the attentional shifting and orientation (Klein, 2000).

Prior knowledge (e.g., memory, prediction, and expectation) also plays a crucial role in human auditory perception, therefore several top-down- and bottom-up-based models integrate the prior knowledge into the data-driven models. Some of them extract acoustic features of the target sound and store them in memory-like modules to mimic humans’ long-term memory function as top-down modulation. Oldoni et al. (2013) combine a self-organized map (SOM) of the acoustic features in the bottom-up processing to continuously learn the saliency and novelty of acoustic features. After training, each SOM unit matches up with a representative sound prototype. For the top-down processing, the IOR and LEGION mechanisms are introduced to shift and select attention, respectively. Xu et al. (2015) propose an Auditory Selection framework with Attention and Memory (ASAM). In this model, there is one speech perceptor extracting the voiceprint of speakers and accumulating the voiceprint in a lifelong-learning memory module during the training phase to be the prior knowledge for the model. Later, the learned voiceprint is used to attend and filter the target speech from the sound input to achieve the top-down and bottom-up interaction. The testing performance showed good robustness and adaptation for both top-down (follow a specific conversation) and bottom-up (capture the salient sound or speech) attention tasks.

Shi et al. (2018) propose the Top-Down Auditory model (TDAA) and use the prediction of the target speaker as the top-down modulation. Their model contributes to the auditory scene analysis with multiple unknown speakers. They adopt the Short-Time Fourier Transformation (STFT) and Bidirectional Long-Short Term Memory (BiLSTM) to predict the number of the speakers. Later, the classifier recurrent neural networks (RNN) separate the most salient speaker and iterate until no more speakers can be separated to avoid repeated prediction. Finally, an attention module is used to separate each speaker’s spectrum from the spectrum mixture. Besides, binaural models are apt to make use of the spatial localization information to address the Cocktail Party problem. For instance, Ma et al. (2018) train DNNs to localize acoustic features in full 360° azimuth angles. After the training phase, the binaural localization with spectral features is used as prior knowledge in the top-down modulation of the model. Their model serves to predict the speech with different localizations under noisy situations with room reverberation. In summary, those top-down and bottom-up interaction models incorporate mechanisms of processing in the human auditory system.
system. They selectively attend or shift attention to the target speech dynamically rather than only focusing on the stream separation, which can be more adaptive to those uncertain and complex auditory scenarios.

5. AUDIOVISUAL CROSSMODAL SELECTIVE ATTENTION

5.1. Behavioral and Neural Mechanisms of Human Crossmodal Selective Attention

In order to survive in an uncertain and multimodal world, humans develop the ability to integrate and discriminate simultaneous signals from multiple sensory modalities, such as vision, audition, tactile, and olfaction. For example, humans can make use of visual cues like lip movement and body gestures to recognize and localize sounds in noisy circumstances. The crossmodal integration ability is beneficial for humans to localize and perceive objects but can also cause ambiguity. Crossmodal conflicts arise when information from different modalities are incompatible with each other and can result in failures of the crossmodal integration and object recognition. To resolve conflicts, selective attention is required to focus on the task-relevant modality information and to ignore the interference from irrelevant modalities (Veen and Carter, 2006). For humans, the capacity for conflict adaptation plays a crucial role in learning and adapting to the environment. When human toddlers detect any conflict between the current environment and their prior knowledge, they will generate curiosity and be motivated to learn new knowledge or rules (Wu and Miao, 2013). Curiosity is also important for the trial and error learning of robots (Haefz et al., 2019). In this subsection, we mainly talk about behavioral and neural mechanisms of selective attention underlying audiovisual crossmodal integration and conflict resolution.

First, how and when does a crossmodal conflict occur? Previous studies proved that humans tend to integrate visual and auditory stimuli with spatial-temporal linkage into the same object (Senkowski et al., 2008). The “Unity assumption” proposes that when humans believe that the multisensory inputs they perceive are generated from the same source, crossmodal integration occurs (e.g., when students think the speech they hear in the lecture room matches the lip movements of the professor, they believe that the speech is from the professor) (Roseboom et al., 2013). Besides, prior knowledge and experience can generate expectation effects to facilitate object recognition during crossmodal integration. Therefore, when the stimuli from different modalities are spatially (e.g., ventriloquism effect; Choe et al., 1975) or temporally incongruent (e.g., double flash illusion; Roseboom et al., 2013) or contrary to our expectations (e.g., see a cat with a “bark” sound), humans perceive crossmodal conflicts. During the early integration stage, selective attention plays the role of capturing the salient visual and auditory stimuli by bottom-up processing. When conflicts are detected, selective attention executes a top-down modulation from higher-level semantic representations according to the internal goal and relevant modalities. The crossmodal information processing is not only a feed-forward process but also contains backward feedback and recurrent processes, which are important to facilitate the primary sensory processing (Talsma et al., 2010; see Figure 5A).

Second, which modality dominates when humans are confronted with audiovisual conflicts? Lots of studies have examined the “ventriloquism effect,” which originally refers to the strong visual bias during the sound localization (Thurlow and Jack, 1973; Choe et al., 1975; Warren et al., 1981). Research findings show that this strong modality bias changes through the lifespan of a human (Sloutsky, 2003). Compared to toddlers, adults are more likely to have visual stimuli preferences (Sloutsky, 2003). Some researchers argue that the ventriloquism effect...
results from an optimal or suboptimal decision-making strategy, especially when unimodal stimuli are blurred. If the auditory stimuli are more reliable than the visual stimuli, an auditory bias occurs as well (Alais and Burr, 2004; Shams and Kim, 2010; Ma, 2012; Roseboom et al., 2013). To sum up, vision in general has a higher spatial resolution than audition, whereas audition has a higher temporal resolution than vision. As the modality appropriateness hypothesis points out, the information from one modality dominates according to the temporal or spatial features of the audiovisual event and the modality with the higher accuracy (Welch and Warren, 1980).

Third, how do humans resolve crossmodal conflicts? In the conflict-monitoring theory, the module of conflict monitoring (CM) is activated when conflicts are detected and passes the signal to the executive control (EC) module to accomplish the task-related conflict resolution by the top-down attentional control (Botvinick et al., 2001). From the previous findings, to perceive crossmodal signals and detect crossmodal conflicts, selective attention plays the role of gating crossmodal coupling between sensory function areas in a modality-general fashion (Eimer and Driver, 2001; Mcdonald et al., 2003; Convento et al., 2018). However, to solve crossmodal conflicts, selective attention inclines toward processing in a modality-specific fashion (Yang et al., 2017; Mengotti et al., 2018).

Except for some specific vision and audition processing brain areas, the superior colliculus (SC) is a crucial brain area with multisensory convergence zones from visual and auditory primary cortices to higher-level multisensory areas. As it is mentioned in section 3.1, the SC also implements selective attention by orienting both covert and overt attention toward the salient stimulus and triggers corresponding motor outputs (e.g., eye movements, saccades) (Wallace et al., 1998; Meredith, 2002; Krauzlis et al., 2013). Besides, the superior temporal sulcus (STS), inferior parietal sulcus (IPS), frontal cortex (including premotor and ACC), and posterior insula are involved in the crossmodal processing (for review see Calvert, 2001; Stein and Stanford, 2008). Within the crossmodal brain functional network, the STS plays the role of linking unimodal representations (Hertz and Amedi, 2014). The parietal lobe is thought to process representations of visual, auditory, and crossmodal spatial attention (Farah et al., 1989). However, when audiovisual inputs are incongruent, crossmodal attenuations or deactivations occur (Kuchinsky et al., 2012). To resolve conflicts, as human fMRI studies have shown, the dorsal anterior cingulate cortex (dACC) is responsible for dealing with conflicts between the current goal and irrelevant distractors. The dACC is positively correlated with attention orientation and interference suppression (Weissman et al., 2004). Song et al. (2017) conducted a mice experiment by using a task with audiovisual conflicts, where audition was required to dominate vision. They found that when the conflict occurred, the co-activation of the primary visual and auditory cortices suppressed the response evoked by vision but maintained the response evoked by audition in the posterior parietal cortex (PPC). Electrophysiological studies have shown the existence of cells that respond to stimulation in more than one modality to accomplish crossmodal integration and conflict resolution. Diehl and Romanski (2014) found that neurons in the ventrolateral prefrontal cortex (VL PFC) of Macaques were bimodal and nonlinear multisensory. When incongruent faces and vocalizations were presented, those neurons showed significant changes with an early suppression and a late enhancement during the stimulus displaying period. Other experimental evidence argues that coherent oscillations across different modality cortices are the key mechanism of the crossmodal interplay (Wang, 2010). An enhancement of the phase locking for the short-latency gamma-band activity (GBA) is found for the attended multisensory stimuli. The early GBA enhancement enables the amplification and integration of crossmodal task-relevant inputs (Senkowski et al., 2008). Incongruent crossmodal inputs cause a stronger gamma-band coherence than congruent inputs suggesting the involvement of gamma oscillations decoupling under crossmodal binding (Misselhorn et al., 2019). Attentional control during the crossmodal integration and conflict resolution is associated with alpha-band effects from the frontoparietal attention network rather than primary sensory cortices. Frontal alpha oscillations are involved in the top-down perceptual regulation; parietal oscillations are involved in the intersensory reorientation (Misselhorn et al., 2019). Reversed to the gamma oscillation patterns, incongruent conditions showed weaker alpha oscillation changes compared to congruent conditions. This gamma-alpha oscillation cycle pattern is proposed to be the information gating mechanism by inhibiting task-irrelevant regions and selectively routing the task-relevant regions (Jensen and Mazaheri, 2010; Bonnefond and Jensen, 2015). In sum, cortical areas that have multimodal convergence zones accomplish crossmodal integration of projections from visual and auditory primary cortices. Neural oscillations coordinate the temporal synchronization between the visual and auditory modality.

5.2. Computational Models Simulating Human Crossmodal Selective Attention

In robotics, crossmodal research focuses mainly on multisensory binding to make robots interact with the environment with higher robustness and accuracy. Compared with unimodal information, crossmodal information is more beneficial to model complex behaviors or achieve high-level functions on artificial systems, such as object detection (Li et al., 2019), scene understanding (Aytar et al., 2017), lip reading (Mroueh et al., 2015; Chung et al., 2017), etc. In psychology, crossmodal research focuses on how crossmodal information helps humans to recognize objects or events by integrating multimodal information and eliminating the crossmodal ambiguity (Calvert, 2001). In computer science, crossmodal research focused on recognizing one modality by using a multimodal dataset or making use of the data from one single modality and retrieve relevant data of other modalities (Skocaj et al., 2012; Wang et al., 2017). However, compared with unimodal, computational modelings based on crossmodal attention remains lacking. In this section, we particularly introduce the undeveloped computational modeling work on selective attention from the audiovisual crossmodal perspective.
Many studies focus on multimodal fusion (Ramachandram and Taylor, 2017), but research about crossmodal selective attention in computer science is limited. Parisi et al. conducted a series of audiovisual crossmodal conflict experiments to explore human selective attention mechanisms in complex scenarios (Parisi et al., 2017, 2018; Fu et al., 2018). During human behavioral tasks, visual and auditory stimuli were presented in an immersive environment. Four loudspeakers were set behind the corresponding positions on a 180-degree screen, where four human-like avatars with visual cues (lip movement or arm movement) were shown. The visual cue and the sound localization could be congruent or incongruent (e.g., the left-most sound with the right-most avatar’s lip movement). During each trial, human participants were asked to determine where the sound was coming from. Participants had to pay attention to the sound localization and suppress the attentional capture by any visual stimuli. Analyses of human behavior results showed that even though arm moving was visually more salient than lip moving, humans had higher error rates of the sound localization when viewing lip movement. This suggests that lip moving might contain more speech or semantic information so it is more difficult to be ignored. Besides, the magnitude of the visual bias was also significant when the incongruent AV stimuli were coming from the two avatars at the extreme right and left sides of the screen. This indicated a wider integration window than other simplified scenes. Based on the bio-inspired cortico-collricular architecture, deep and self-organizing neural networks consisting of visual and auditory neuron layers and crossmodal neuron layers were used to learn crossmodal integration and selective attention (see Figure 5B). In this way, human-like responses were modeled and embedded in an iCub robot.

The work above shows that computational models can simulate human selective attention on audiovisual sound localization and semantic association. Due to the limited resources and sensory modules, the future exploration of modeling and simulating the attention module is desirable in crossmodal robotics. Besides, selective attention mechanisms can boost the applicability and accuracy of robots in real human-robot interaction scenarios. Robots can select more reliable modalities and reduce distraction and errors.

### 6. CONCLUDING REMARKS AND OUTSTANDING QUESTIONS

The current review summarizes experimental findings, theories, and model approaches of audiovisual unimodal and crossmodal selective attention from psychology, neuroscience, and computer science perspective. Currently, psychologists and neural scientists are working toward computational modeling, standardizing, and replication. In parallel, computer scientists are trying to design and make agent systems more intelligent with higher-level cognitive functions, meta-learning abilities, and lower learning costs. Some advantages, unresolved problems, and future directions of collaborative research in psychology, neuroscience, and computer science are summarized as follows:

#### 6.1. How Psychology, Neuroscience, and Computer Science Benefit From Each Other

One the one hand, findings and methods from psychology and neuroscience can interpret and improve models’ performance (Hohman et al., 2018). For instance, representational similarity analysis (RSA) is nowadays also used to compare the responses recorded in fMRIs and artificial systems like deep learning CNNs. RSA analyzes the similarity of fMRI responses and brain representations by a set of stimuli (Kriegeskorte et al., 2008). Dwivedi and Roig (2019) found that RSA shows good performance on transfer learning and task taxonomy by computing correlations between the models on certain tasks. On the other hand, the-state-of-the-art approaches offer tools to analyze big data of neural findings. For example, the SyConn framework used deep CNNs and random forest classifiers to accelerate data analyses on animal brains to compute the synaptic wiring of brain areas (Dorkenwald et al., 2017). Another application of computational modeling is examining theories and interpreting mechanisms in human behaviors or neural responses (O’Reilly, 2006). The key idea is to examine crucial cognitive function in hidden layers of the modal. Models can be built to simulate normal behaviors and then mimic the “damage” by changing parameters of sub-units. If the “damage” causes similar abnormal behaviors as psychiatric patients do, the changed units may be the corresponding mechanisms to the behaviors. For instance, Wang and Fan (2007) collected human behavioral data by the ANT and used leabra (local, error-driven, and associative, biologically realistic algorithm) model (O’Reilly, 1998) to explore the potential interaction between each functional network (alerting, orienting, and executive control). Their model successfully simulated healthy human behavior. After changing one parameter of the executive control module, their model could simulate the behavior of schizophrenic patients, suggesting the crucial role of executive control.

#### 6.2. Limits Remain in Current Interdisciplinary Research

Even though we have reviewed and summarized a number of findings from psychology and computer science, lots of unsolved issues of attention processing remain to be disclosed. The simulation work of crossmodal attention and conflict processing is insufficient on robots. Besides, the problem of perceptual constancy has not been deeply addressed in computer science. For humans, it is easy to recognize one object from different perspectives, such as finding an open door in a dim room. Moreover, humans can transfer the intrinsic knowledge to learn and infer new objects or concepts with a small number of learning samples. However, artificial intelligent systems cannot reach humans’ performance yet. For example, even though the scale-invariant feature transform (SIFT) algorithm (Lowe, 1999) can extract features from variant shapes of the same object, it cannot recognize the variant objects when only colors exist without any structural patterns. Current deep learning approaches like the VGG net (Simonyan and Zisserman, 2015) has shown better
performance on object recognition than traditional approaches. However, such deep networks rely on the training dataset and need substantial computational resources.

6.3. Future Directions for Interdisciplinary Research

There is a lot of potential for psychologists and computer scientists to work together to investigate both human cognition and intelligent systems. On the one hand, psychologists can focus on designing paradigms to diagnose and remedy shortages of current models to improve the model accuracy. Besides, neural studies are still needed to understand human brain mechanisms better. It will be insightful to develop bio-inspired computational models with a better interpretability. On the other hand, for computer science, enhancing the complexity of models to increase the adaptability and flexibility to the environment is required. At last, to balance the computational complexity and biological plausibility is also crucial, because humans’ behavioral patterns are limited by their capacity and energy load, even though the properties of machines will keep improving. In summary, deepening the understanding of each processing mechanism rather than only describing phenomena is the direction for research from both sides to endeavor.

REFERENCES

Ahveninen, J., Jääskeläinen, I. P., Raji, T., Bonmassar, G., Devore, S., Hamalainen, M., et al. (2006). Task-modulated “what” and “where” pathways in human auditory cortex. Proc. Natl. Acad. Sci. U.S.A. 103, 14608–14613. doi: 10.1073/pnas.0510480103

Alain, C., Arnott, S. R., Hevenor, S., Graham, S., and Grady, C. L. (2001). “What” and “where” in the human auditory system. Proc. Natl. Acad. Sci. U.S.A. 98, 12301–12306. doi: 10.1073/pnas.211209098

Alais, D., and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. Curr. Biol. 14, 257–262. doi: 10.1016/j.cub.2004.01.029

Anderson, B. A., Laurent, P. A., and Yantis, S. (2011). Value-driven attentional capture. Proc. Natl. Acad. Sci. U.S.A. 108, 10367–10371. doi: 10.1073/pnas.1104047108

Awh, E., Belopolsky, A. V., and Theeuwes, J. (2012). Top-down versus bottom-up attentional control: a failed theoretical dichotomy. Trends Cogn. Sci. 16, 437–443. doi: 10.1016/j.tics.2012.06.010

Aytar, Y., Castrejon, L., Vondrick, C., Pirsiavash, H., and Torralba, A. (2017). Cross-modal scene networks. IEEE Trans. Patt. Anal. Mach. Intell. 40, 2303–2314. doi: 10.1109/TPAMI.2017.2753232

Ba, J., Mnih, V., and Kavukcuoglu, K. (2014). “Multiple object recognition with visual attention,” in International Conference on Learning Representations (Banff, AB).

Bacon, W. F., and Egeth, H. E. (1994). Overriding stimulus-driven attentional capture. Percept. Psychophys. 53, 485–496. doi: 10.3758/BF03205306

Baddley, A., Hitch, G., and Bower, G. (1974). Recent advances in learning and motivation. Work Mem. 8, 647–667.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). “Neural machine translation by jointly learning to align and translate,” in International Conference on Learning Representations (Banff, AB).

Bai, S., and An, S. (2018). A survey on automatic image caption generation. Neurocomputing 311, 291–304. doi: 10.1016/j.neucom.2018.05.080

Barbey, A. K. (2018). Network neuroscience theory of human intelligence. Trends Cogn. Sci. 22, 8–20. doi: 10.1016/j.tics.2017.10.001

Bee, M. A., and Micheyl, C. (2008). The cocktail party problem: what is it? How can it be solved? And why should animal behaviorists study it? J. Comp. Psychol. 122, 235–251. doi: 10.1037/0735-7036.122.3.235

Benes, F. M. (2000). Emerging principles of altered neural circuitry in schizophrenia. Brain Res. Rev. 31, 251–269. doi: 10.1016/S0165-0173(99)00041-7

Bizley, J. K., and Cohen, Y. E. (2013). The what, where and how of auditory-object perception. Nat. Rev. Neurosci. 14, 693–707. doi: 10.1038/nrn3565

Bonnehof, M., and Jensen, O. (2015). Gamma activity coupled to alpha phase as a mechanism for top-down controlled gating. PLoS ONE 10:e0128667. doi: 10.1371/journal.pone.0128667

Borji, A., and Itti, L. (2012). State-of-the-art in visual attention modeling. IEEE Trans. Patt. Anal. Mach. Intell. 35, 185–207. doi: 10.1109/TPAMI.2012.89

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., and Cohen, J. D. (2001). Conflict monitoring and cognitive control. Psychol. Rev. 108, 624–652. doi: 10.1037/0033-295X.108.3.624

Bregman, A. S. (1994). Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA: The MIT Press.

Broadbent, D. E. (2013). Perception and Communication. Amsterdam: Elsevier.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. J. Acoust. Soc. Am. 109, 1101–1109. doi: 10.1121/1.1345696

Bullmore, E., and Sporns, O. (2012). The economy of brain network organization. Nat. Rev. Neurosci. 13, 336–349. doi: 10.1038/nrn3214

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. Cereb. Cortex 11, 1110–1123. doi: 10.1093/cercor/11.12.1110

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. 25, 975–979. doi: 10.1121/1.1907229

Choe, C. S., Welch, R. B., Gilford, R. M., and Juola, J. F. (1975). The “ventriloquist effect”: visual dominance or response bias? Percept. Psychophys. 18, 55–60. doi: 10.3758/BF03199367

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). “Attention-based models for speech recognition,” in Advances in Neural Information Processing Systems (Montreal, QC), 577–585.

AUTHOR CONTRIBUTIONS

SW and XL contributed to the conception and organization of the manuscript. DF wrote the first draft of the manuscript. CW, KY, MK, WN, PB, HW, XL, and SW authors contributed to manuscript reading and revision and approved the submitted version.

FUNDING

This work was supported by National Natural Science Foundation of China (NSFC, No. 61621136008), the German Research Foundation (DFG) under project Transregio Crossmodal Learning (TRR 169), and CAS-DAAD joint fellowship.

ACKNOWLEDGMENTS

We especially appreciate feedback and support from the Knowledge Technology Group in Hamburg and the Carelab in Beijing. We thank Katja Kösters for proofreading the manuscript and improving the language clarity. We also particularly thank Dr. German I. Parisi, Ge Gao, Zhenghan Li, Honghui Xu, and Antonio Andriella for the fruitful discussions.
Togo, F., Lange, G., Natelson, B. H., and Quigley, K. S. (2015). Attention network test: assessment of cognitive function in chronic fatigue syndrome. J. Neuropsychol. 9, 1–9. doi: 10.1011/jnp.12030

Treisman, A., and Gormican, S. (1988). Feature analysis in early vision: evidence from search asymmetries. Psychol. Rev. 95, 15–48. doi: 10.1037/0033-295X.95.1.15

Uddin, L. Q. (2015). Salience processing and insular cortical function and dysfunction. Nat. Rev. Neurosci. 16:55. doi: 10.1038/nrn3857

Uddin, L. Q., and Menon, V. (2009). The anterior insula in autism: under-connected and under-examined. Neurobiol. Behav. Rev. 33, 1198–1203. doi: 10.1016/j.neubiorev.2009.06.002

Urbanek, C., Weinges-Evers, N., Bellmann-Strohli, J., Bock, M., Dörr, J., Hahn, E., et al. (2010). Attention network test reveals alerting network dysfunction in multiple sclerosis. Multiple Scler. J. 16, 93–99. doi: 10.1177/1352458609350308

Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., and Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve visual spatial search. J. Exp. Psychol.-Hum. Percept. Perform. 34:1053. doi: 10.1037/0096-1523.34.5.1053

Van der Burg, E., Olivers, C. N., Bronkhorst, A. W., and Theeuwes, J. (2009). Poke and pop: tactile-visual synchrony increases visual saliency. Neurosci. Lett. 460, 60–64. doi: 10.1016/j.neulet.2008.11.002

VanRullen, R. (2003). Visual saliency and spike timing in the ventral visual pathway. J. Physiol. Paris 97, 365–377. doi: 10.1016/j.jphysparis.2003.09.010

Varela, F., Lachaux, J.-P., Rodriguez, E., and Martinerie, J. (2001). The brainweb: phase synchronization and large-scale integration. Nat. Rev. Neurosci. 2:229. doi: 10.1038/35067550

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). “Attention is all you need,” in Proceedings of International Conference on Machine Learning (Lille), 2048–2057.

Wei, L., Qu, Z., Feng, Y., and Zhu, Y. (2015). Distinct dynamic computational principles of cortical rhythms in cognition. Physiol. Rev. 99, 1159–1208. doi: 10.1152/physrev.00035.2015

Wen, Y., Huang, M., Zhu, X., and Zhao, L. (2016). “Attention-based LSTM for aspect-level sentiment classification,” in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (Austin, TX), 606–615. doi: 10.18653/v1/D16-1058

Welch, R. B., and Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. Psychol. Bull. 88, 638–667. doi: 10.1037/0033-2909.88.3.638

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., et al. (2015). “Show, attend and tell: neural image caption generation with visual attention,” in Proceedings of International Conference on Machine Learning (Lille), 2048–2057.

Yang, G., Nan, W., Zheng, Y., Wu, H., Li, Q., and Liu, X. (2017). Distinct cognitive control mechanisms as revealed by modality-specific conflict adaptation effects. J. Exp. Psychol. Hum. Percept. Perform. 43:807. doi: 10.1037/xhp0000351

Yantis, S., and Jonides, J. (1984). Abrupt visual onsets and selective attention: evidence from visual search. J. Exp. Psychol. Hum. Percept. Perform. 16:601. doi: 10.1037/0096-1523.15.3.601

Yao, X., Han, J., Zhang, D., and Nie, F. (2017). Revisiting co-saliency detection: a novel approach based on two-stage multi-view spectral rotation co-clustering. IEEE Trans. Image Process. 26, 3196–3209. doi: 10.1109/TIP.2017.2694222

Yin, P., Fritz, J. B., and Shamma, S. A. (2014). Rapid spectrotemporal plasticity in primary auditory cortex during behavioral performance. J. Neurosci. 34, 4396–4408. doi: 10.1523/JNEUROSCI.2799-13.2014

Zhang, X., and Gong, Q. (2019). Frequency-following responses to complex tones comply with these terms.

Zouridakis, G., Simos, P. G., and Papanicolaou, A. C. (1998). Multiple bilaterally connected and under-examined.

Zhou, S., and Smolensky, P. (2017). “Attention is all you need,” in Proceedings of International Conference on Machine Learning (Lille), 2048–2057.

Copyright © 2020 Fu, Weber, Yang, Kerzel, Nan, Barros, Wu, Liu and Wermter. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.