frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 19 May 2020
doi: 10.3389/fncom.2020.00015

Conscious Multisensory Integration: Introducing a Universal Contextual Field in Biological and Deep Artificial Neural Networks

Ahsan Adeel[1,2]*

[1] Oxford Computational Neuroscience, Nuffield Department of Surgical Sciences, John Radcliffe Hospital, University of Oxford, Oxford, United Kingdom, [2] School of Mathematics and Computer Science, University of Wolverhampton, Wolverhampton, United Kingdom

Conscious awareness plays a major role in human cognition and adaptive behavior, though its function in multisensory integration is not yet fully understood, hence, questions remain: How does the brain integrate the incoming multisensory signals with respect to different external environments? How are the roles of these multisensory signals defined to adhere to the anticipated behavioral-constraint of the environment? This work seeks to articulate a novel theory on conscious multisensory integration (CMI) that addresses the aforementioned research challenges. Specifically, the well-established contextual field (CF) in pyramidal cells and coherent infomax theory (Kay et al., 1998; Kay and Phillips, 2011) is split into two functionally distinctive integrated input fields: local contextual field (LCF) and universal contextual field (UCF). LCF defines the modulatory sensory signal coming from some other parts of the brain (in principle from anywhere in space-time) and UCF defines the outside environment and anticipated behavior (based on past learning and reasoning). Both LCF and UCF are integrated with the receptive field (RF) to develop a new class of contextually-adaptive neuron (CAN), which adapts to changing environments. The proposed theory is evaluated using human contextual audio-visual (AV) speech modeling. Simulation results provide new insights into contextual modulation and selective multisensory information amplification/suppression. The central hypothesis reviewed here suggests that the pyramidal cell, in addition to the classical excitatory and inhibitory signals, receives LCF and UCF inputs. The UCF (as a steering force or tuner) plays a decisive role in precisely selecting whether to amplify/suppress the transmission of relevant/irrelevant feedforward signals, without changing the content e.g., which information is worth paying more attention to? This, as opposed to, unconditional excitatory and inhibitory activity in existing deep neural networks (DNNs), is called conditional amplification/suppression.

Keywords: universal contextual field, pyramidal cell, multisensory integration, coherent infomax neuron, contextually-adaptive neuron, deep neural network, audio-visual speech processing

OPEN ACCESS

Edited by:
Sliman J. Bensmaia,
University of Chicago, United States

Reviewed by:
Radwa Khalil,
Jacobs University Bremen, Germany
Jorge F. Mejias,
University of Amsterdam, Netherlands

***Correspondence:**
Ahsan Adeel
ahsan.adeel@deepci.org

Received: *22 September 2019*
Accepted: *07 February 2020*
Published: *19 May 2020*

Citation:
Adeel A (2020) Conscious Multisensory Integration: Introducing a Universal Contextual Field in Biological and Deep Artificial Neural Networks. Front. Comput. Neurosci. 14:15. doi: 10.3389/fncom.2020.00015

1. INTRODUCTION

What is conscious awareness? Think of a well-trained and experienced car driver who automatically identifies and follows the traffic protocols in different surrounding environments (e.g., street, highway, city centre) by simply interpreting the visual scenes directly (such as buildings, school etc.). Similarly, imagine a car with slightly defective parking sensors that sometimes

miscalculates the distance to the nearest object. In this case, the audio input is ambiguous and the driver cannot fully rely on parking sensors for precise maneuvering decisions, e.g., while reversing the car. To tackle this problem, the driver automatically starts utilizing visual cues to leverage the complementary strengths of both ambiguous sound (reversing-beeps) and visuals for optimized decision making. These are a few examples of conscious awareness, where the external environment helps establishing the anticipated behavior and the corresponding optimal roles of incoming multisensory signals.

Nonetheless, it raises crucial questions: How does it happen in the brain? How do the incoming sensory signals (such as vision and sound) integrate with respect to the situation? How does a neuron originate a precise control command complying with the anticipated behavioral-constraint of the environment? Certainly, defining the context and its relevant features knowing when a change in context has taken place are challenging problems in modeling human behavior (Gonzalez et al., 2008). It is also claimed in the literature that context could be of infinite dimensions but humans have a unique capability of correlating the significant context and set its boundaries intuitively (Gonzalez et al., 2008). However, once the context is identified, it is relatively easy to utilize and set its bounds to more precisely define the search space for the selection of best possible decision (Gonzalez et al., 2008).

A simple example of contextual modulation is shown in **Figure 1**. It can be seen that the ambiguous RF input (in the top row) is interpreted as "B" or "13" depending on the LCF (i.e., "A," "C," "12," and "14") and UCF (i.e., knowledge of English alphabets and numeral system). Similarly, it is observed that in noisy environments (e.g., a busy restaurant, bar, cocktail party), human brain naturally utilizes other modalities (such as lips, body language, facial expressions) to perceive speech or the conveyed message [i.e., speech-in-noise (SIN) perception] (Sumby and Pollack, 1954; McGurk and MacDonald, 1976; Summerfield, 1979; Patterson and Werker, 2003). This multimodal nature of speech is well-established in the literature; it is understood how speech is produced by the vibration of vocal folds and configuration of the articulatory organs. The developed AV speech processing models are depicted in **Figure 2**. Two distinctive input variables RF and LCF are defining the incoming sensory inputs (i.e., sound and vision), whereas the UCF input is defining three different surrounding environments: Restaurant, Cafe, and Home. In any environment, multisensory information streams are available, but their optimal integration depends on the outside environment. For example, in a busy cafe and restaurant environment (multi-talker speech perception), the processor utilizes other modalities (i.e., lips as LCF) to disambiguate the noisy speech, whereas in the Home scenario (with little or zero noise), LCF has a Null role.

Hence, coordination and specialization are necessary to produce coherent thoughts, percepts, and actions, which are well-adapted to different situations and long-term goals (Phillips et al., 2015). However, the understanding of specialization and coordination is still a major issue within the cognitive and neurosciences. The hypothesis reviewed in Phillips et al. (2015) suggests that this is mostly achieved by a widely distributed process of contextual modulation, which amplifies and suppresses the transmission of signals that are relevant and irrelevant to current circumstances, respectively. Nevertheless, selective modulation (amplification/attenuation) of incoming multisensory information with respect to the outside world is poorly understood. In addition, not much progress has been made on the use of conscious awareness and contextual modulation to show enhanced processing, learning, and reasoning. In this research article, the aforementioned interesting observations are discussed and a new perspective in terms of CMI with some future research directions is comprehensively presented. The rest of the paper is organized as follows: section 2 discusses the conceptual foundation and motivation that leads to the development of a CAN model. Section 3 presents the CAN and contextually-adaptive neural network (CANN) structures. In sections 4 and 5, the proposed theory is utilized for AV speech processing. Finally, conclusion and future research directions are presented in sections 6 and 7, respectively.

2. MOTIVATION AND CONTRIBUTION

For a long time, it was believed that the consciousness depends on neurons firing and synchronization at certain frequency bands. Massimini et al. (2005) suggested that consciousness is not critically dependent on them, but rather on the ability of the brain to integrate multisensory information. This brain ability depends on the effective connectivity[1] among functionally specialized regions of the thalamocortical system. At a granular level, evidence gathered in the literature suggests that the multisensory interaction emerges at the primary cortical level (Stein and Stanford, 2008; Stein et al., 2009). The divisive/multiplicative gain modulations are widely spread in mammalian neocortex with an indication of amplification or attenuation via contextual modulation (Galletti and Battaglini, 1989; Salinas and Sejnowski, 2001; Phillips et al., 2018).

Scientists have presented several models and empirical results on the role of contextual modulation to disambiguate the ambiguous input (Kay and Phillips, 1997; Kay et al., 1998; Phillips, 2001; Phillips and Silverstein, 2013). For example, in Kay et al. (1998), the contextual modulation was demonstrated using a simple edge detection problem to reveal its effectiveness in recognizing specific patterns with noisy RF input. It was shown how surrounding regions (CF) in different parallel streams helped detecting the edge within any particular region and played a significant role in combating noisy input. This idea is called a coherent infomax theory or coherent infomax neuron (Kay et al., 2017; Lizier et al., 2018).

The physiological studies in Phillips and Singer (1997) have suggested that biological neurons, in addition to the classical excitatory and inhibitory signals, do receive contextual inputs. These contextual inputs possibly fulfill the gain-controlling RC role (Fox and Daw, 1992). The authors in Kepecs and Raghavachari (2002) used a two-compartment model of

[1]Effective connectivity is the ability of neuronal groups to causally affect other neuronal groups within a system.

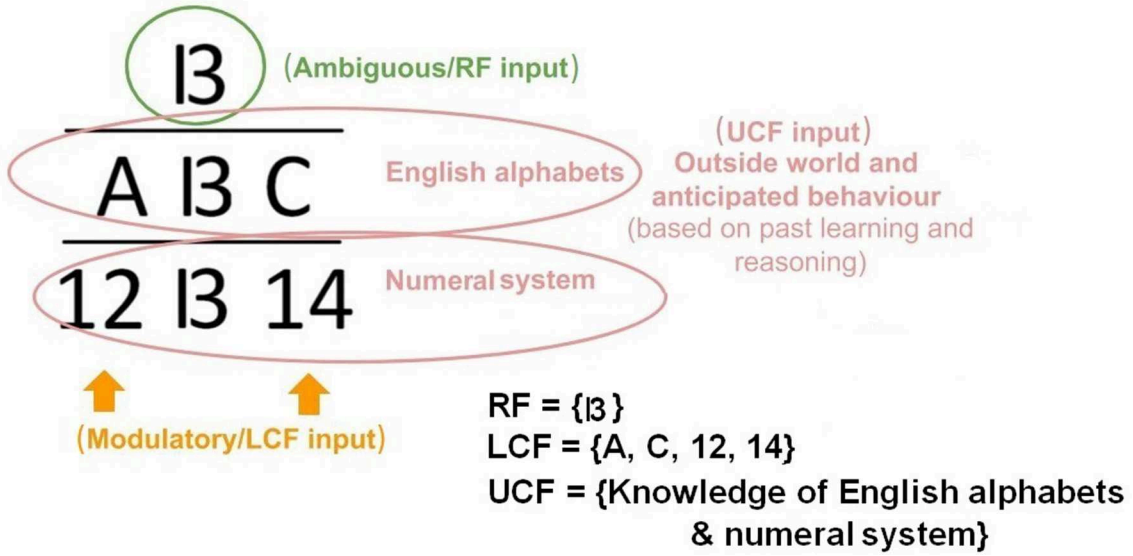

FIGURE 1 | Ambiguous decision making and contextual modulation.

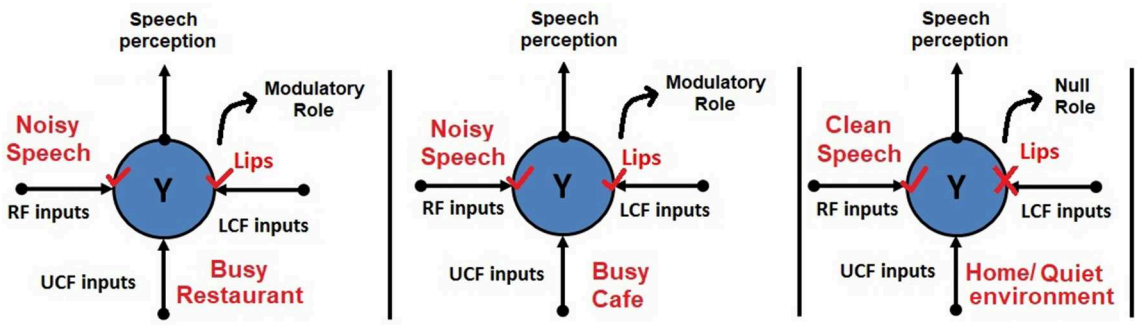

FIGURE 2 | Human AV speech modeling in three different environments. Please note that the role of LCF changes with respect to the outside environment (UCF). For example, in the first two environments, LCF has a modulatory role, whereas in the third environment, it has a Null role.

pyramidal neurons to capture the spatial extent of neuronal morphology. Their study simulated three neurons, each receiving the same α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA), representing the informational input i.e., a word "green." The three neurons also received distinct contextual input via NMDA receptors, representing specific noun groups: objects, people and fruits. It is to be noted that the word "green," when expressed with a contextual input, varies in meaning, e.g., a color green or an unripe fruit. The simulation results showed that even though each neuron received the same strong AMPA input, their firing was uncorrelated and context-dependent.

An overlay of a coherent infomax neural processor on layer 5 pyramidal cells is shown in **Figure 3A** (Wibral et al., 2017), highlighting potential parallels to existing physiological mechanisms. In two sites of integration, one is at the soma and the other at the top of the apical trunk. The driving excitatory (R_e) or inhibitory (R_i) signals arrive via basal and perisomatic synapses, whereas the modulatory excitatory (C_e) or inhibitory (C_i) signals arrive via synapses on the tuft dendrites at the top of the apical trunk. Na^+ at the somatic integration site initiates sodium spikes that backpropagate up to the apical trunk. Ca^{2+} at the apical integration site initiates calcium spikes, which amplify the neural response (Phillips et al., 2015).

In light of the aforementioned literature, in this paper, the contextual AV speech processing is used to demonstrate contextual modulation. Specifically, the CF in coherent infomax theory (Kay et al., 1998; Kay and Phillips, 2011) is split into two fields: LCF and UCF. LCF defines the modulatory sensory signal coming from some other parts of the brain (in principle from anywhere in space-time) and UCF defines the outside environment and anticipated behavior (based on past learning and reasoning). Both LCF and UCF are integrated with the RF to develop a new class of contextually-adaptive neuron, which adapts to changing situations (shown in **Figure 3B**). For evaluation and comparative analysis, two distinctive multimodal multistreams (lip movements as LCF and noisy speech as

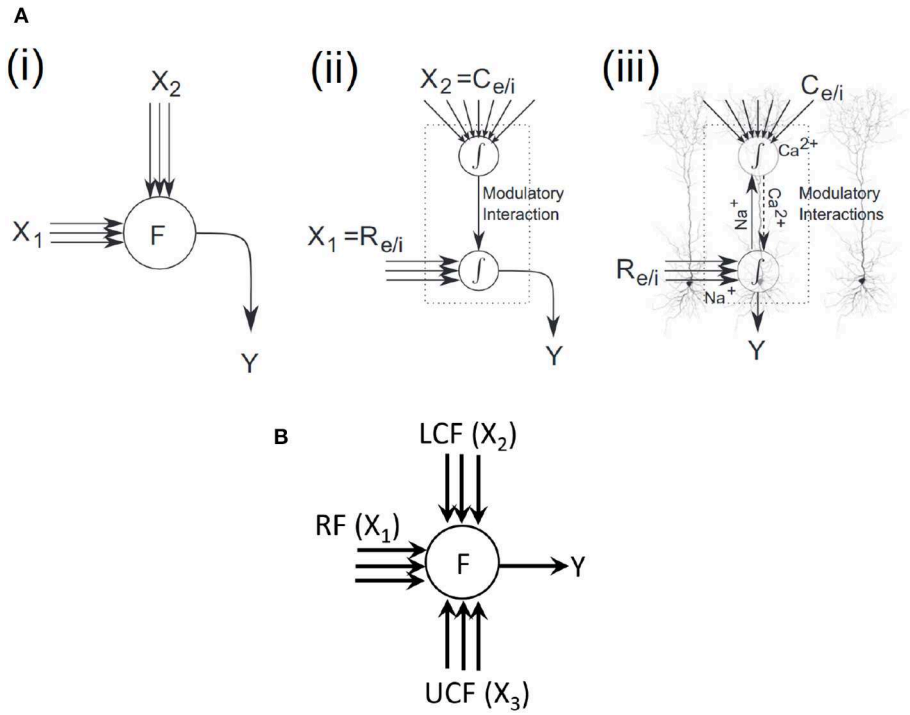

FIGURE 3 | Coherent infomax vs. contextually-adaptive neural processor. **(A)** Coherent infomax neural processor (Phillips et al., 2015; Wibral et al., 2017): (i) with multidimensional inputs *X1;X2*, and output *Y*. (ii) with local weighted summation of inputs: X1 receptive field [excitatory (R_e) or inhibitory (R_i)] and X2 contextual field [excitatory (C_e) or inhibitory (C_i)]. (iii) overlay on a layer 5 pyramidal cells. **(B)** Proposed CAN.

RF) are used to study the role of LCF in SIN perception (ranging from a very noisy environment to almost zero noise). Furthermore, going beyond the theory of coherent infomax, UCF is introduced as a fourth new dimension to represent the outside environment and anticipated behavior. Its effectiveness is shown in terms of enhanced learning and processing, using three distinctive multimodal multistreams (lip movements as LCF, noisy speech as RF, and outside environment/anticipated behavior as UCF).

3. CONTEXTUALLY-ADAPTIVE NEURON (CAN)

The proposed CAN is presented in **Figure 4**. The output of the neuron depends on three functionally distinctive integrated input variables: driving (RF), modulatory (LCF), and UCF. The RF is defining the ambiguous sensory signal, LCF is defining the modulatory sensory signal coming from other parts of the brain, and UCF is defining the outside world and anticipated behavior. The interaction among RF, LCF, and UCF is shown in **Figure 5A**. The output is denoted by the random variable Y, whereas X, Z, and U represent RF, LCF, and UCF, respectively. In CANN, the CAN in one stream is connected to all other CANs in the neighboring stream of the same layer as shown in **Figure 5B**. This is achieved through shared connections among the neurons that guide learning and processing with respect to local and universal contexts.

3.1. Mathematical Modeling

The CAN (Y) in CANN interacts by exchanging the excitatory and inhibitory spikes probabilistically (in the form of bipolar signal trains). In steady state, the stochastic spiking behavior of the network has a "product form" property (product of firing rates and transition probabilities) which defines a state probability distribution with easily solvable non-linear network equations. The firing from neuron y to succeeding neuron w in the network is according to the Poisson process, represented by the synaptic weights $w^+_{yw} = r_y[P^+_{yx} + P^+_{yz} + P^+_{yu}]$ and $w^-_{yw} = r_y[P^-_{yx} + P^-_{yz} + P^-_{yu}]$, where $P^+_{yx}, P^+_{yz}, P^+_{yu}$ and $P^-_{yx}, P^-_{yz}, P^-_{yu}$ represent the probabilities of excitatory and inhibitory RF, LCF, and UCF signals, respectively. The term r_y represents the firing rate of the CAN. The terms w^+_{yx}, w^+_{yz}, w^+_{yu} and w^-_{yx}, w^-_{yz}, w^-_{yu} represent the RF, LCF, and UCF synaptic weights (i.e., the rates of positive and negative signal transmission) that network learns through the process of learning or training. In the network, CAN receives exogenous signals positive/negative from the inside (within the network) or outside world, according to Poisson arrival streams of rates Λ_x, λ_x, respectively. The potential (Y) of the CAN represents its state that increases/decreases with respect to an incoming signal coming from the inside or outside world. The proposed neural structure is implemented using G-networks that possess a product-form asymptotic solution (Gelenbe, 1993a).

The CAN in firing state transmits an impulse to neuron w with a Poisson rate (r_y) and probability $P^+(y, w)$ or $P^-(y, w)$ depending on the incoming signal being excitatory or inhibitory.

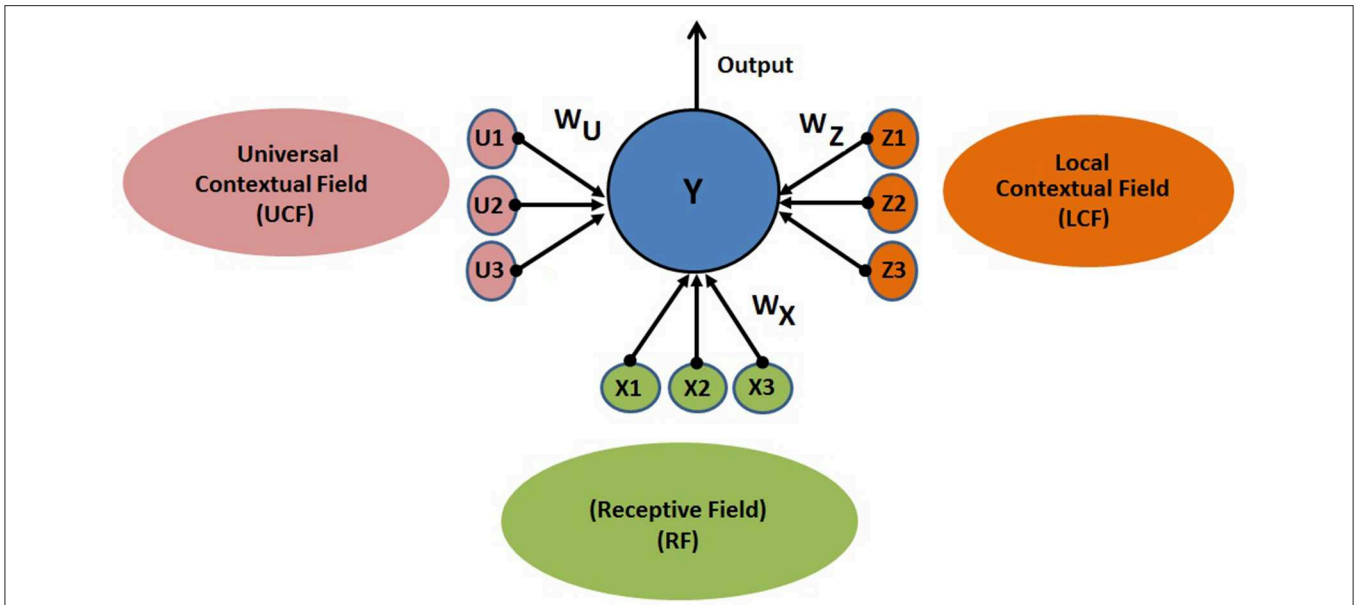

FIGURE 4 | CAN structure: the output depends on three functionally distinctive integrated input variables: driving (RF), modulatory (LCF), and UCF. The RF is defining the ambiguous sensory signal, LCF is defining the modulatory sensory signal coming from other parts of the brain, and UCF is defining the outside world and anticipated behavior. W_X, W_Z, and W_U are representing the receptive, local contextual, and universal contextual field connections, respectively.

The transmitted signal can also leave the network and go outside the world with probability $d(y)$ such that:

$$d(y) + \sum_{x=1}^{N}[P^+(y,x) + P^-(y,x)] + \sum_{z=1}^{N}[P^+(y,z) + P^-(y,z)] + \sum_{u=1}^{N}[P^+(y,u) + P^-(y,u)] = 1 \quad (1)$$

Where,

$$w^+(y,w) = r_y[P^+(y,x) + P^+(y,z) + P^+(y,u)] \geq 0,$$
$$w^-(y,w) = r_y[P^-(y,x) + P^-(y,z) + P^-(y,u)] \geq 0 \quad (2)$$

The firing rate of CAN can be written as:

$$r(y) = (1 - d(y))^{-1}(\sum_{x=1}^{N}[w^+(y,x) + w^-(y,x)] + \sum_{z=1}^{N}[w^+(y,z) + w^-(y,z)] + \sum_{u=1}^{N}[w^+(y,u) + w^-(y,u)]) \quad (3)$$

If $Y(t)$ is the potential of CAN then in n number of neurons, vector $\overline{Y(t)} = (y_1(t), y_2(t), ..., y_n(t))$ can be modeled as a continuous-time Markov process. The stationary joint probability of the network is given as:

$$\lim_{n\to\infty} P(\overline{Y(t)}) = y_1(t), y_2(t), ..., y_n(t) = \prod_{y=1}^{n}(1 - q_y)q_y^{ny},$$
$$q_y = \frac{Q_Y^+}{r_y + Q_Y^-} \quad (4)$$

where Q_Y^+ and Q_Y^- are the average rates of +ive and -ive signals at the CAN (y), given as:

$$Q_Y^+ = \sum_{x=1}^{N} q_x w^+(y,x) + \sum_{z=1}^{N} q_z w^+(y,z) + \sum_{u=1}^{N} q_u w^+(y,u) \quad (5)$$

$$Q_Y^- = \sum_{x=1}^{N} q_x w^-(y,x) + \sum_{z=1}^{N} q_z w^-(y,z) + \sum_{u=1}^{N} q_u w^-(y,u) \quad (6)$$

The probability that CAN (Y) is excited can be written as:

$$q_y = \frac{\sum_{x=1}^{N} q_x w^+(y,x) + \sum_{z=1}^{N} q_z w^+(y,z) + \sum_{u=1}^{N} q_u w^+(y,u)}{[W_W^+ + W_W^-] + \sum_{w=1}^{N} q_x w^-(y,x) + \sum_{z=1}^{N} q_z w^-(y,z) + \sum_{u=1}^{N} q_u w^-(y,u)} \quad (7)$$

where $w^+(y,x)$, $w^-(y,x)$, $w^+(y,z)$, $w^-(y,z)$ $w^+(y,u)$, $w^+(y,u)$ are the positive and negative RF, LCF, and UCF weights. W_W^+ and W_W^- are the positive and negative weights between CAN and

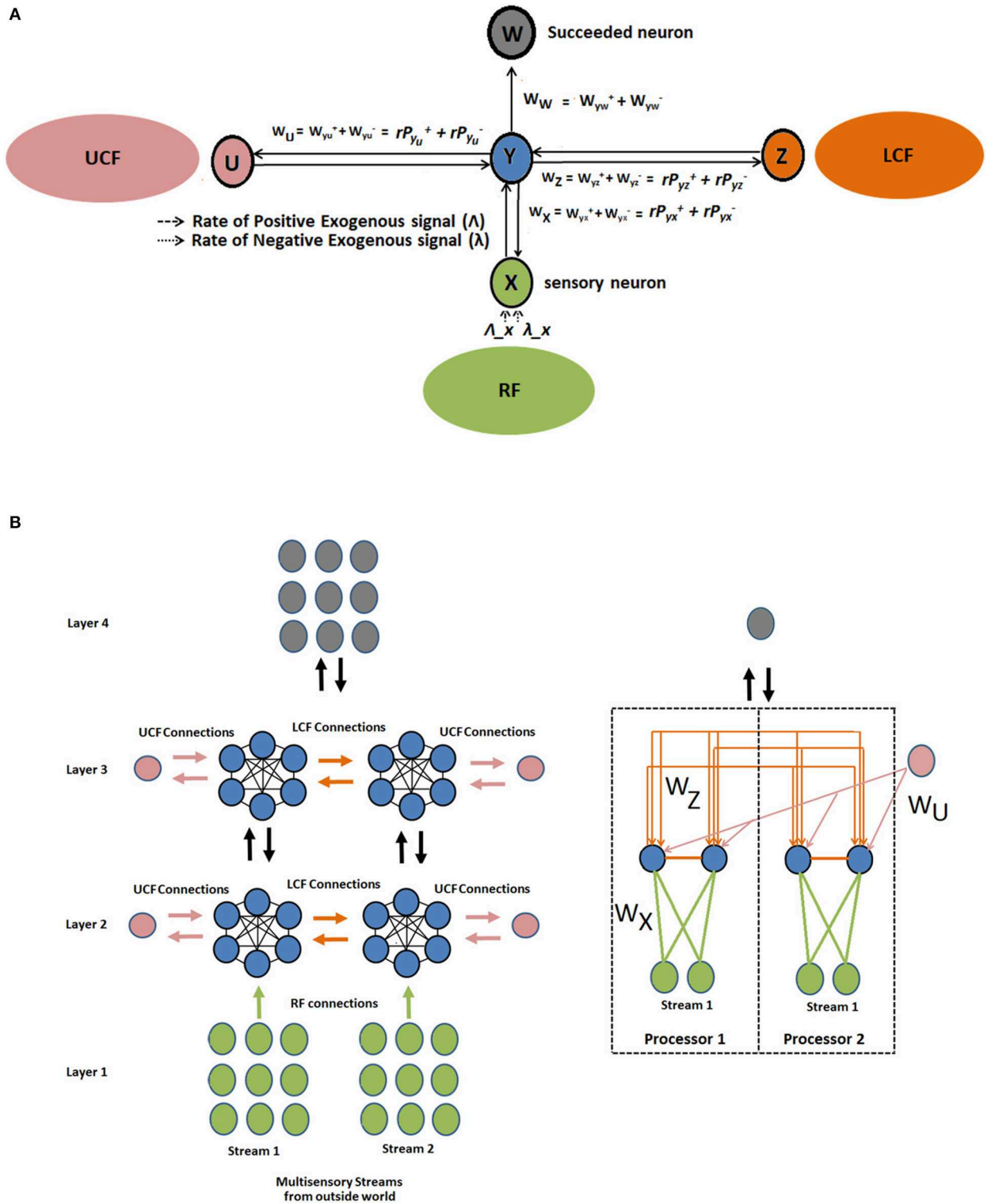

FIGURE 5 | Interaction among RF, LCF, and UCF in CAN and CANN. **(A)** CAN: The filtering rules (precise information integration) are enforced by the positive and negative synaptic weights associated with each input field. **(B)** CANN: multilayered multiunit network of similar CANs, where the CAN in one stream is connected to all other CANs in neighboring streams of the same layer. The figure on the right is providing detailed information about connections using two RF and LCF units in each processor, with one UCF and one output unit each.

succeeded neuron w. For training and weights update, state-of-the-art gradient descent algorithm is used (Gelenbe, 1993b). The RF input (q_x) is given as:

$$q_x = \frac{Q_x^+}{[w(x,y)^+ + w(x,y)^-] + Q_x^-} \tag{8}$$

$$Q_x^+ = \Lambda_x + \sum_{v=1}^{N} q_v w^+(x,v) \tag{9}$$

$$Q_x^- = \lambda_x + \sum_{v=1}^{N} q_v w^-(x,v) \tag{10}$$

where q_v is the potential of the preceding neuron v and q_u and q_z are potentials of the incoming UCF and LCF neurons, respectively. It is to be noted that $w(x,y)^+$ and $w(y,x)^+$ are different.

3.2. Information Decomposition

A Venn diagram of the information theoretic measures for distinctive integrated input variables is depicted in **Figure 6**, where RF, LCF, and UCF are represented by the green, orange, and grayish pink ellipses, respectively. The output (Y) is represented by the blue ellipse. In information processing equations, the output is denoted by the random variable Y, whereas RF, LCF, and UCF are represented by X, Z, and U, respectively.

The mutual information shared between random variables X (RF) and Y (output) can be written as (Kay and Phillips, 2011):

$$I(X;\, Y) = H(X) - H(X|Y) \tag{11}$$

Where, H(X) is the Shannon entropy associated with the distribution of X and $H(X|Y)$ is the Shannon entropy associated with the conditional distribution of X given Y. It is defined as the information contained in X but not in Y (Kay and Phillips, 2011). It is assumed that the mutual information is always non-negative when random variables are stochastically independent (Kay and Phillips, 2011). Since we are dealing with four random variables, the conditional mutual information can be written as:

$$I(X;\, Y|Z, U) = H(Y|Z, U) - H(Y|X, Z, U) \tag{12}$$

This is the conditional mutual information shared between X and Y, having observed Z and U. It is defined as the information shared between X and Y but not shared with Z and U.

The four-way mutual information shared among four random variables X, Y, Z, and U can be defined as:

$$\begin{aligned} I(X;\, Y;\, Z;\, U) &= I(X;\, Y) - I(X;\, Y|Z, U) \\ &= I(X;\, Z) - I(X;\, Z|Y, U) = \\ I(X;\, U) - I(X;\, U|Y, Z) &= I(Y;\, Z) - I(Y;\, Z|X, U) \\ &= I(Y;\, U) - I(Y;\, U|X, Z) \end{aligned} \tag{13}$$

If the four-way mutual information is positive, Shannon entropy associated with the distribution of Y can be defined as (Kay and Phillips, 2011):

$$\begin{aligned} H(Y) = I(Y;\, X;\, Z;\, U) + I(Y;\, X|Z, U) + I(Y;\, Z|X, U) \\ + I(Y;\, U|X, Z) + H(Y;\, X|Z, U) \end{aligned} \tag{14}$$

In case the random variables are discrete, the integrals are replaced by summations, and the probability mass function can be written as (Kay and Phillips, 2011):

$$H(Y) = -\int p(y) \log p(y) dy \tag{15}$$

$$H(Y|X) = -\int\int p(y|x) \log p(y|x) p(x) dy dx \tag{16}$$

$$\begin{aligned} H(Y|X, Z) = -\int\int\int p(y|x, z) \\ \log p(y|x, z) p(x, z) dy dx dz \end{aligned} \tag{17}$$

$$\begin{aligned} H(Y|X, Z, U) = -\int\int\int\int p(y|x, z, u) \\ \log p(y|x, z, u) p(x, z, u) dy dx dz du \end{aligned} \tag{18}$$

The objective function to be maximized can be defined as:

$$\begin{aligned} F = \phi_0 I(Y;\, X;\, Z;\, U) + \phi_1 I(Y;\, X|Z, U) \\ + \phi_2 I(Y;\, Z|X, U) + \phi_3 I(Y;\, U|X, Z) \\ + \phi_4 H(Y;\, X|Z, U) \end{aligned} \tag{19}$$

$I(Y;\, X|Z, U)$ is the information that the output shares with the RF (X) and is not contained in the LCF and UCF units. $I(Y;\, Z|X, U)$ is the information that the output shares with the LCF and not contained in the RF and UCF units. $I(Y;\, U|X, Z)$

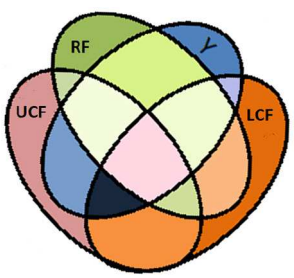

FIGURE 6 | Venn diagram of information theoretic measures for distinctive integrated input variables RF, LCF, and UCF represented by the green ellipse, orange ellipse, and grayish pink ellipse, respectively. The output (Y) is represent by the blue ellipse. The UCF (U) and associated $H(U|X, Y, Z)$ is interpreted as the information contained in U but not in X, Y, and Z. The output (Y) and associated $H(Y|X, Z, U)$ is interpreted as the information contained in Y but not in X, Z, and U. The LCF (Z) and associated $H(Z|X, Y, U)$ is interpreted as the information contained in Z but not in X, Y, and U. The RF (X) and associated $H(X|Y, Z, U)$ is interpreted as the information contained in X but not in Y, Z, and U.

is the information that the output shares with the UCF and not contained in the RF and LCF units.

The values of $\phi's$ are tunable within the range [−1, 1]. Different ϕ values allow investigating specific mutual/shared information, such that:

$$f(x) = \begin{cases} F = I(Y;X), & \text{if } \phi_1 = 1, \phi_2 = \phi_3 = \phi_4 = 0 \\ F = I(Y;Z), & \text{if } \phi_2 = 1, \phi_1 = \phi_3 = \phi_4 = 0 \\ F = I(Y;U), & \text{if } \phi_3 = 1, \phi_1 = \phi_2 = \phi_4 = 0 \\ I(Y;X;Z;U), & \text{otherwise} \end{cases}$$

4. CASE STUDY: HUMAN BEHAVIORAL MODELING/AV SPEECH PROCESSING

Human speech recognition in a noisy environment is known to be dependent upon both aural and visual cues, which are combined by sophisticated multi-level integration strategies to improve intelligibility (Adeel et al., 2019b). The correlation between the visible properties of articulatory organs (e.g., lips, teeth, tongue) and speech reception has been previously shown in numerous behavioral studies (Sumby and Pollack, 1954; McGurk and MacDonald, 1976; Summerfield, 1979; Patterson and Werker, 2003). Therefore, clear visibility of some articulatory organs could be effectively utilized to extract a clean speech signal out of a noisy audio signal. The proposed CMI theory is evaluated using human AV speech modeling. The developed AV models are illustrated in section 1 and **Figure 2**.

4.1. Audio-Visual Corpus and Feature Extraction

For contextual AV speech modeling, the AV ChiME3 corpus is developed by mixing the clean Grid videos (Cooke et al., 2006) with the ChiME3 noises (Barker et al., 2015) [cafe, street junction, public transport (bus), pedestrian area] for signal-to-noise ratio (SNRs) ranging from −12 to 12 dB (Adeel et al., 2019b). The pre-processing includes sentence alignment and incorporation of prior visual frames. Sentence alignment is performed to remove the silence time from the video and prevent the model from learning redundant or insignificant information. Prior multiple visual frames are used to incorporate temporal information to improve mapping between visual and audio features. The Grid corpus comprises 34 speakers, each speaker reciting 1,000 sentences. Out of 34 speakers, a subset of 5 speakers is selected (two white females, two white males, and one black male) with a total of 900 command sentences each. The subset fairly ensures the speaker independence criteria (Adeel et al., 2019b). A summary of the acquired visual dataset is presented in **Tables 1**, **2**, where the full and aligned sentences, total number of sentences, used sentences, and removed sentences are clearly defined (Adeel et al., 2019b).

For audio features, log filter-bank (FB) vectors are used. The input audio signal is sampled at 50 kHz and segmented into N 16 ms frames with 800 samples per frame and 62.5% increment rate. Afterwards, a hamming window and Fourier transformation is applied to produce the 2,048-bin power spectrum. Finally, a 23-dimensional log-FB is applied, followed by the logarithmic compression to produce the 23-D log-FB signal (Adeel et al., 2019b).

TABLE 1 | Used grid corpus sentences (Adeel et al., 2019b).

			Full sentences		Aligned sentences	
Speaker ID	**Grid ID**	**No. of sentences**	**Removed**	**Used**	**Removed**	**Used**
Speaker 1	S1	1,000	11	989	11	989
Speaker 2	S15	1,000	164	836	164	836
Speaker 3	S26	1,000	16	984	71	929
Speaker 4	S6	1,000	9	991	9	991
Speaker 5	S7	1,000	11	989	11	989

TABLE 2 | Summary of the train, test, and validation sentences (Adeel et al., 2019b).

Speakers	**Train**	**Validation**	**Test**	**Total**
1	692	99	198	989
2	585	84	167	836
3	650	93	186	929
4	693	99	199	991
5	692	99	198	989
All	3,312	474	948	4,734

The visual features are extracted from the Grid Corpus videos recorded at 25 fps using a 2D-DCT based standard and a widely used visual feature extraction method. Firstly, the video files are processed to extract a sequence of individual frames. Secondly, a Viola-Jones lip detector (Viola and Jones, 2001) is used to identify the lip-region by defining the Region-of-Interest (ROI) in terms of a bounding box. Object detection is performed using Haar feature-based cascade classifiers. The method is based on machine learning where cascade function is trained with positive and negative images. Finally, the object tracker (Ross et al., 2008) is used to track the lip regions across the sequence of frames. The visual extraction procedure produced a set of corner points for each frame, where lip regions are then extracted by cropping the raw image. In addition, to ensure good lip tracking, each sentence is manually validated by inspecting a few frames from each sentence. The aim of manual validation is to delete those sentences in which lip regions are not correctly identified (Abel et al., 2016; Adeel et al., 2019b).

5. EXPERIMENTS

Signal processing in the cerebral cortex is expected to comprise a common multipurpose algorithm that produces widely distributed but coherent and relevant activity patterns (Kay and Phillips, 2011). The coherent infomax exhibits specification of such algorithm. According to the theory of coherent infomax, local processors are able to combine reliable signal coding because of the existence of two classes of synaptic connections: driving connections (RF) and contextual connections (CF). The authors in Kay and Phillips (2011) made the biological relevance of this theory and showed that the coherent infomax is consistent

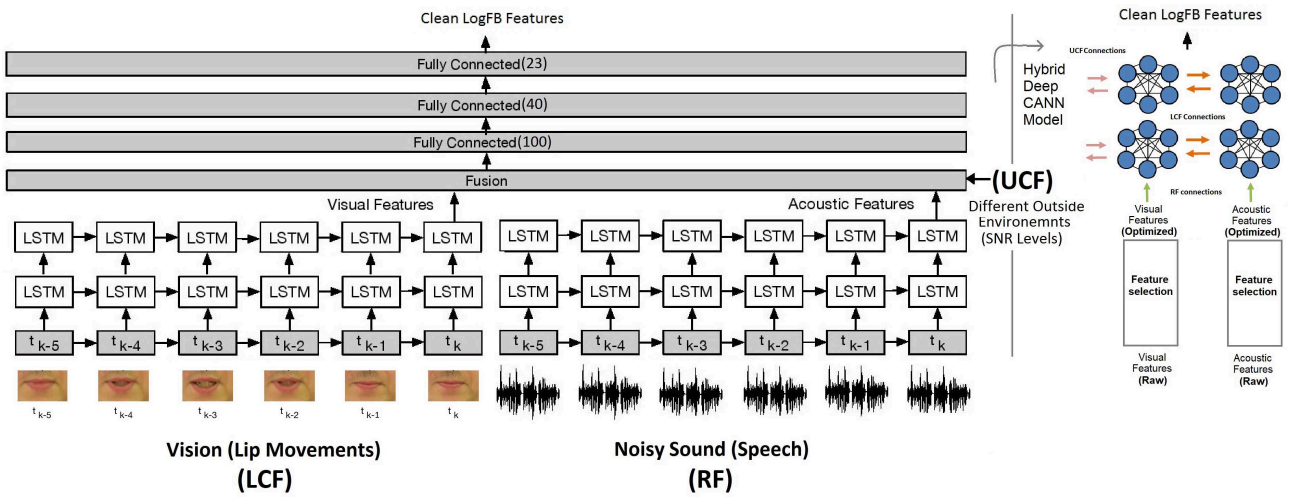

FIGURE 7 | State-of-the-art network level AV deep LSTM model with 1000 LSTM and 162 MLP cells (noisy audio to clean audio mapping): CMI at the network level (Adeel et al., 2019a). The figure on the right depicts the hybrid deep LSTM-CANN model.

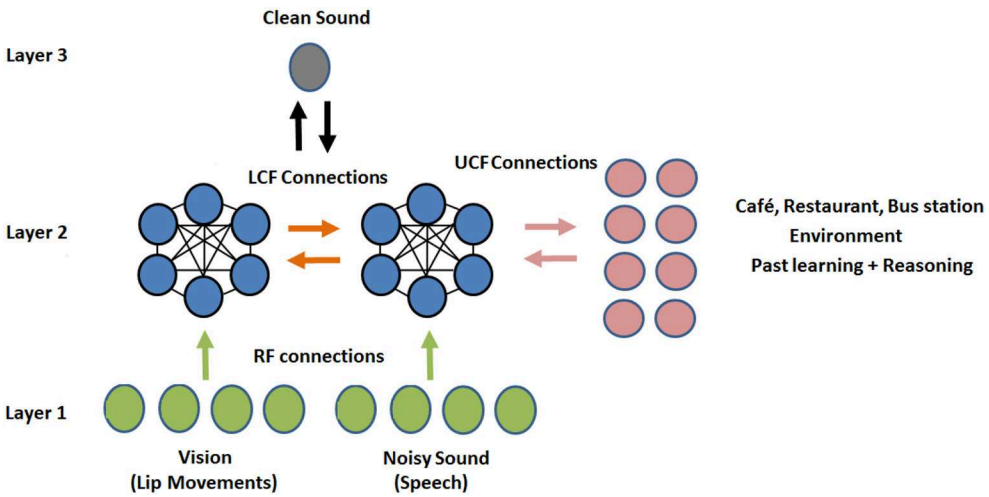

FIGURE 8 | Proposed novel shallow CANN: CMI at the neural level. Three streams, three layered multisensory multiunit network of several similar CANs (where the CAN in one stream is connected to all other CANs in the neighboring streams). The CANN has 12 CANs, 4 RF neurons, 4 LCF neurons, 8 UCF neurons, and 1 output neuron, in total 29 neurons.

with a particular Bayesian interpretation for the contextual guidance of learning and processing. However, this theory was evaluated using a simple edge detection problem to demonstrate the role of contextual modulation in improving feature detection with noisy inputs (Kay et al., 1998). The authors showed that how surrounding regions in different parallel streams (via contextual modulation) helped detecting the edges within any particular region and played a modulatory role in combating noisy input. More details including different properties of the coherent infomax are comprehensively presented in Kay and Phillips (1997), Kay et al. (1998), Phillips (2001), Phillips and Silverstein (2013), Kay et al. (2017), and Phillips et al. (2018).

Going beyond a simple edge detection problem, in this subsection, we demonstrate how parallel streams constituting visual information play a modulatory role to disambiguate the noisy speech signal. For this, a hybrid deep LSTM and CANN models are developed for network and neural level multisensory integrations, shown in **Figure 7** (Adeel et al., 2019a) and **Figure 8**, respectively. In **Figure 7**, the model on the right depicts the hybrid deep LSTM-CANN, which is a part of our ongoing work. Both LSTM and CANN models are trained with the AV ChiME3 dataset for SNRs ranging form −12 to 12 dB. Noisy audio and visual features of time instance $t_k, t_{k-1}, ..., t_{k-5}$ are feeded into LSTM and CANN models. The aim is to map noisy audio to clean audio features. The first LSTM layer has 250 cells, which encoded the input and passed its hidden state to the second LSTM layer, which has 300 cells. Finally, the optimized latent features from both LSTM models are fused

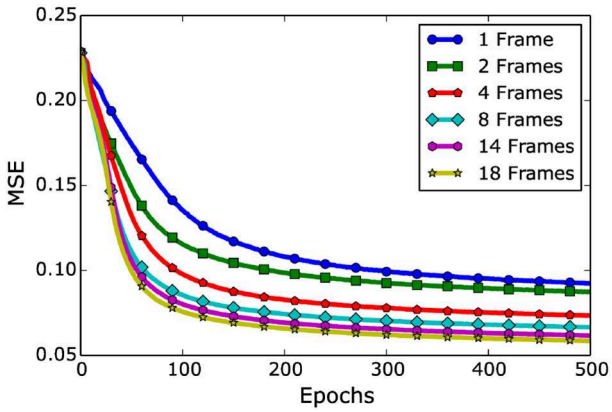

FIGURE 9 | Single stream (RF only): Network level visual only (lip-reading driven) deep learning performance. The figure presents an overall behavior of an LSTM model when different number of visual frames are added. Please note the improvement in learning upon prior visual frames integration (Adeel et al., 2019b).

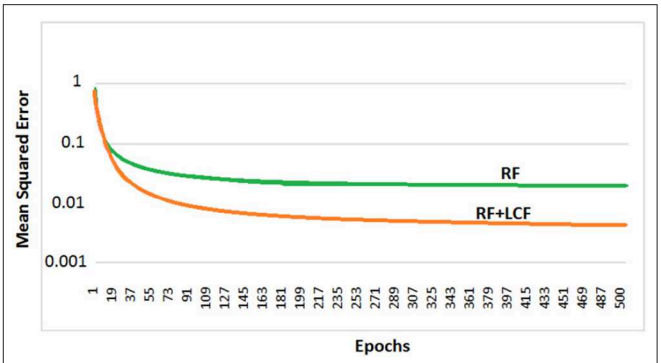

FIGURE 10 | Parallel multistreams (RF + LCF): learning and processing results. It is to be noted that the integration of LCF significantly improved learning.

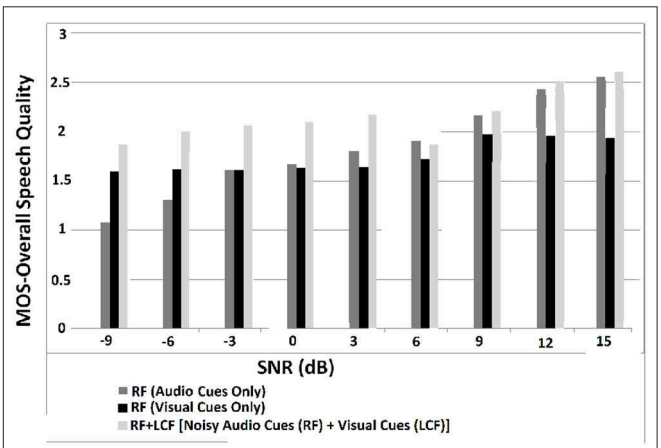

FIGURE 11 | MOS for overall speech perception—these results are obtained through subjective listening tests, conducted in terms of MOS with self-reported normal-hearing listeners. The listeners were presented with a single stimulus (i.e., enhanced speech only) and were asked to rate the re-constructed speech on a scale of 1–5. The five rating choices were: (5) Excellent (when the listener feels an unnoticeable difference compared to the target clean speech), (4) Good (perceptible but not annoying), (3) Fair (slightly annoying), (2) Poor (annoying), and (1) Bad (very annoying).

using two dense layers. Specifically, the optimal features extracted from each LSTM network are concatenated into a single vector. The concatenated vector is then feeded into a fully connected multilayered perceptron (MLP) network. The MLP network comprises three layers with 100 and 40 ReLU neurons in the first two layers and 22 linear neurons in the last layer. The UCF is integrated in the second last layer. The training procedure for CANN is comprehensively presented in section 5.3.1. Both deep LSTM and CANN architectures are trained with the objective to minimize the mean squared error (MSE) between the predicted and the actual clean audio features. The MSE (20) between the estimated audio logFB features and clean audio features is minimized using the stochastic gradient decent algorithm and the RMSProp optimizer. RMSprop is an adaptive learning rate optimizer which divides the learning rate by the moving average of the magnitudes of recent gradients to make learning more efficient. Moreover, to reduce overfitting, dropout (0.20) was applied after every LSTM layer. The MSE cost function $C(a_{estimated}, a_{clean})$ can be written as (Adeel et al., 2019b):

$$C(a_{\text{estimated}}, a_{\text{clean}}) = \sum_{i=1}^{n} 0.5(a_{\text{estimated}}(i) - a_{\text{clean}}(i))^2 \qquad (20)$$

where $a_{estimated}$ and a_{clean} are the estimated and clean audio features, respectively.

5.1. Single Stream: RF Only

The deep LSTM model is trained only with visual cues (RF only) considering multiple prior frames (ranging from 1 visual frame to 18 prior visual frames). The simulation results are shown in **Figure 9** (Adeel et al., 2019b). The training is performed with six different aligned datasets (i.e., 1, 2, 4, 8, 14, and 18 prior visual frames). It can be seen that by moving from 1 visual frame to 18 visual frames, a significant performance improvement could be achieved. The LSTM model with 1 visual frame achieved the MSE of 0.092, whereas with 18 visual frames, the model achieved the least MSE of 0.058. The LSTM model exploited the temporal information effectively and showed consistent reduction in MSE. This is mainly because of its inherent recurrent architectural property and the ability of retaining state over longer time spans using cell gates.

5.2. Parallel Multistreams: RF + LCF

In this experiment, the deep LSTM model is feeded with noisy audio cues (as RF) and visual cues (as LCF). The training results of AV model (RF + LCF) are depicted in **Figure 10**, where the improvement in learning and processing due to LCF integration is evident (Adeel et al., 2019a). The speech perception results in terms of speech quality are shown in **Figure 11**. The used speech enhancement framework is out of the scope of this paper and is comprehensively presented in Adeel et al. (2019b). It can be

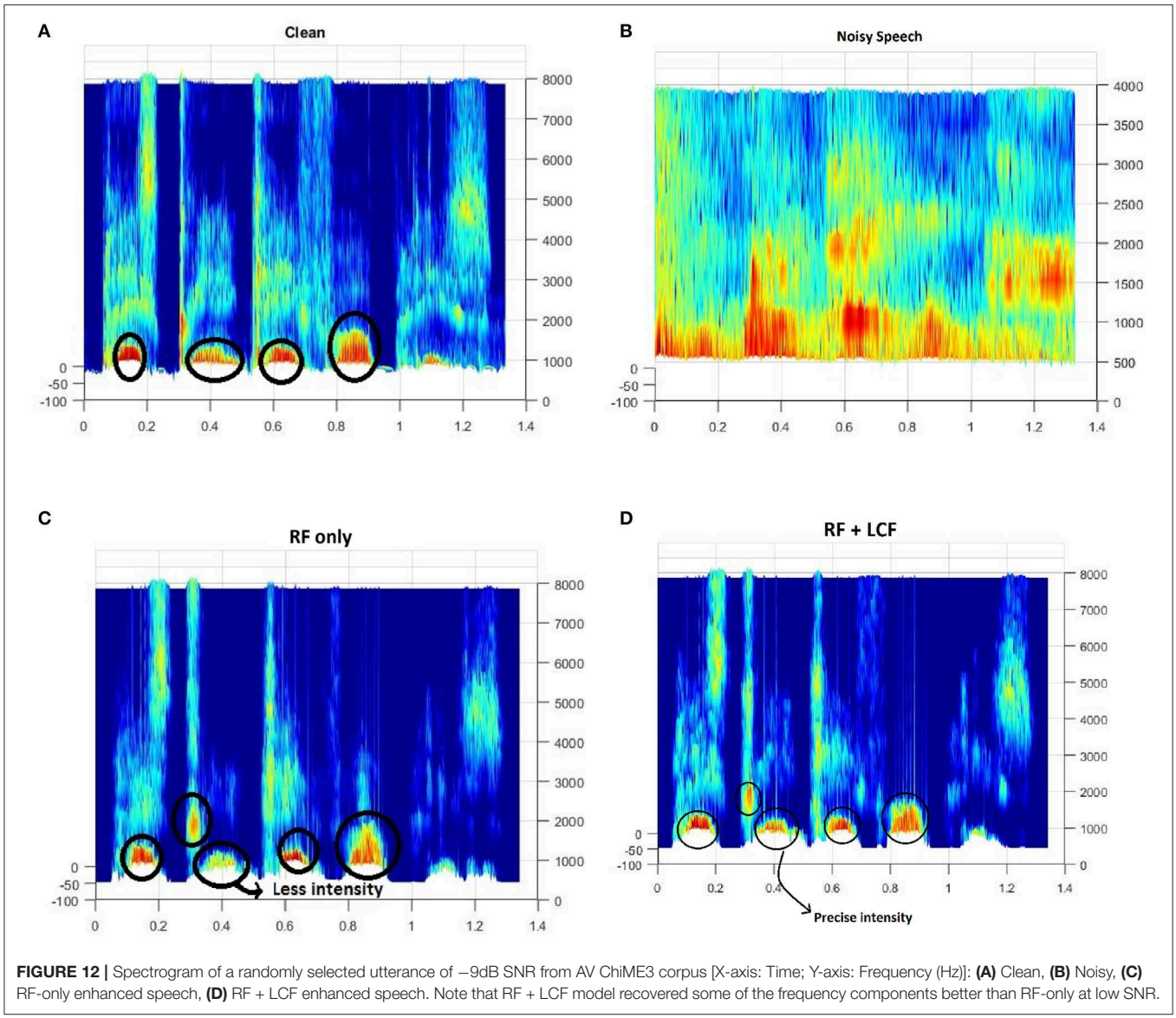

FIGURE 12 | Spectrogram of a randomly selected utterance of −9dB SNR from AV ChiME3 corpus [X-axis: Time; Y-axis: Frequency (Hz)]: **(A)** Clean, **(B)** Noisy, **(C)** RF-only enhanced speech, **(D)** RF + LCF enhanced speech. Note that RF + LCF model recovered some of the frequency components better than RF-only at low SNR.

seen that at high level of background noise (e.g., busy restaurant), visual-only cues are outperforming audio-only cues. In contrast, at low level of background noise (high SNR), audio-only cues are outperforming visual-only cues. It shows that visual cues are fairly less effective for speech enhancement at low or zero background noise which is analogous to human audio-visual speech processing. However, AV (RF + LCF) model outperforms both audio-only and visual-only models in all situations (at both low and high SNRs). The AV model is leveraging the complementary strengths of both audio and visual cues. At high background noise, LCF is acting as a modulatory signal and helping the model to disambiguate the noisy audio speech signal. At low level of background noise, the role of LCF starts decreasing [eventually reaches to Null (hypothetically)—ceiling effect]. This phenomenon is more clear in **Figure 12**, where the spectrogram of a randomly selected utterance of −9 dB SNR is depicted. However, for in-depth and neural level analysis, the CANN is modeled and trained in section 5.3.1 to enable better quantification of this amplification and suppression process.

5.3. Beyond Coherent Infomax: RF + LCF + UCF

So far, it is seen how LCF could play a modulatory or null role upon changing the context (outside environment). However, contextual identification and transition (from one context to another) are two difficult problems. Given any desired human behavior to be modeled, a set of appropriate contexts associated with the anticipated behaviors and actions could be identified and grouped together to develop a computationally efficient model (given a broader understanding of the task in hand) (Gonzalez et al., 2008). In this subsection, the deep

LSTM model is trained with three distinctive multimodal multistreams: lip movements as LCF, noisy speech as RF, and the outside environment/anticipated behavior as UCF. For contextual information (UCF integration), five dynamic real-world commercially-motivated scenarios are considered: cafe, restaurant, public transport, pedestrian area, and home. Please note that a specific SNR range defines a particular environment (UCF), represented by a unique pattern. The training results are presented in **Figure 13** where a significant improvement in learning is evident. Hence, given a broader understanding of the task in hand (acquired through incoming sensory signals (having a high correlation to the external world), specific situation, and associated anticipated behavior), an enhanced learning and optimized decision making could be achieved.

5.3.1. CANN: RF + LCF + UCF

So far, the end-to-end multimodal deep learning models have demonstrated CMI at the network level. However, the underlying neural processing in deep learning models is elusive and it is difficult to analyze the precise information processing. For example, in case of deep LSTM driven AV processing, it is difficult to quantify the selective amplification or suppression of multisensory AV information at different levels. To address these problems, the proposed novel AV-CANN model (shown in **Figure 8**) is trained using AV ChiME3 corpus. For training, the deep problem was transformed into a shallow problem. Specifically, the evaluated shallow CANN model predicts one coefficient at a time (i.e., coefficient by coefficient prediction). The data samples include 2D-logFB (speech features) and 2D-DCT (visual features) coefficients for 1,000 utterances from Grid and ChiME3 Corpora (Speaker 1 of the Grid). The number of clean logFB audio features are 22 × 205,712. The combined noisy logFB audio features are 22 × 205,712 (for −12, −9, −6, −3, 0, 3, 6, 9, and 12 dB SNRs). Similarly, the DCT visual features are 25 × 205,712 in total.

In AV CANN model, the filter bank (audio cues) and DCT (visual cues) coefficients are represented as signals coming from the outside world according to Poisson arrival streams of rates (Λ_x, λ_x). These inputs are converted into average rate of positive and negative signals, given by Equations (9) and (10). Specifically, a set of successive inputs is denoted as X = ($x^{(1)}$...., $x^{(K)}$), Where, x(k) = ($\Lambda_x^{(k)}$, $\Lambda_x^{(k)}$) are pairs of excitation and inhibition signals entering each neuron from the outside world.

Figure 14 depicts the prediction of clean logFB coefficients, where it can be seen that RF + LCF + UCF model outperformed both RF + LCF and RF-only models, achieving MSE of 0.051, 0.064, and 0.072, respectively. It is also worth mentioning that the shallow CANN with only 29 spiking neurons performed comparably to deep LSTM unimodal network (RF-only). In conjunction with the coherent infomax theory, the enhanced learning in CANN is due to a widely distributed and shared activity pattern. The CANN discovered and exploited the associative relations between the features extracted within each of the RF, LCF, and UCF streams.

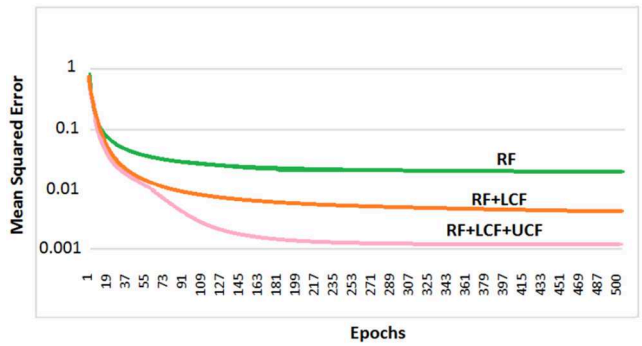

FIGURE 13 | Deep LSTM learning and processing results: it is to be noted that RF + LCF + UCF model outperforms both RF-only and RF + LCF models.

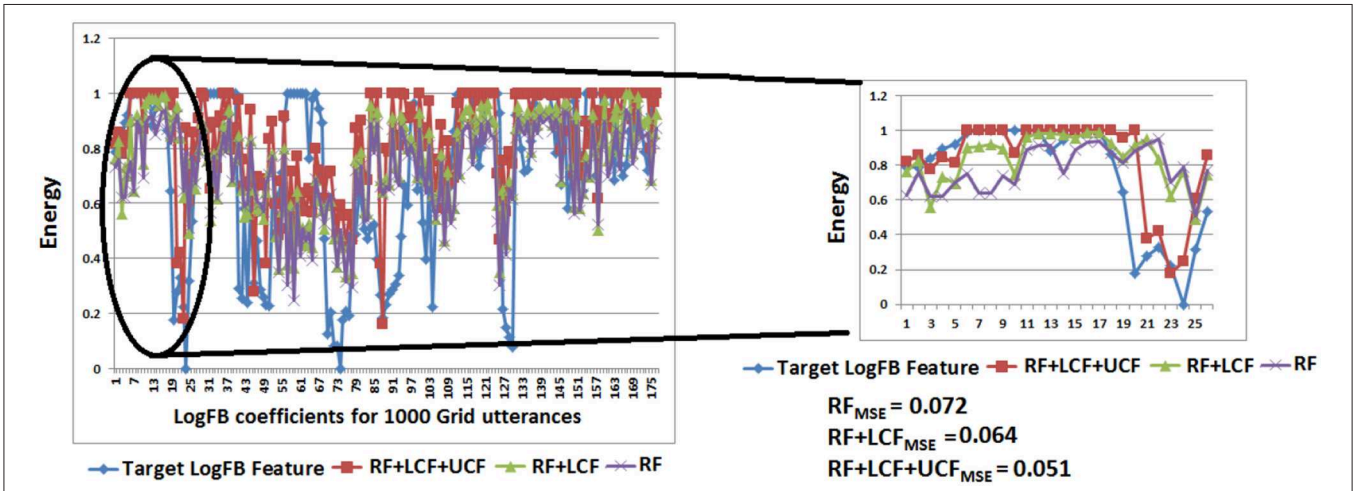

FIGURE 14 | CANN learning and processing results: clean audio features prediction with A-only (RF), AV (RF + LCF), and AV with UCF models. It is to be noted that RF + LCF + UCF model outperforms both RF-only and RF + LCF models.

6. DISCUSSION AND CONCLUSIONS

It is worth mentioning that the author has not claimed to know the origin of consciousness, instead, proposed a theory on its possible function in multisensory integration. A two-compartment neuron with distinct somatic (RF) and apical (CF) zones of integration is well-established (Larkum et al., 2009; Kay and Phillips, 2011; Larkum, 2013; Larkum and Phillips, 2016) and supported for effective learning in deep networks (Lillicrap et al., 2020). However, the apical input (CF), coming from the feedback and lateral connections, is far more diverse with far greater implications for ongoing learning and processing in the brain. CMI theory emphasizes the importance of understanding and defining the roles of different kinds of contexts in pyramidal cells. Thus, it puts forward the idea of dissecting CF into LCF and UCF, to better understand the amplification and suppression of relevant and irrelevant signals, with respect to different external environments and anticipated behaviors.

Preliminary results shed light on selective amplification/attenuation of AV signals. It is shown that in different environmental conditions (represented as UCF), roles of audio and visual cues change, e.g., in high-level of background noise, visual cues (as LCF) modulate the noisy audio cues (RF), whereas, in low-level of background noise, LCF becomes relatively less effective, with no role (hypothetically) in zero background noise. Furthermore, in terms of enhanced learning and processing, the parallel three-stream (RF + LCF + UCF) deep neural network model out performs the parallel two-stream (RF + LCF) and single-stream (RF-only) models. Similar results are obtained with a shallow contextually-adaptive neural network (CANN), which also enables quantification of multiway mutual/shared information at the neural level. The integration of RF, LCF, and UCF guides learning and processing while enables the network to explore and exploit the associative relations between the features extracted within different fields for optimized decision making.

These findings suggest that the pyramidal cell, in addition to the classical excitatory and inhibitory signals, receives the LCF and UCF inputs. The UCF (as a steering force or tuner) helps pyramidal cells in precisely selecting the relevant or useful feedforward signals from overwhelming available information and deciding whether to amplify/suppress their transmission e.g., which information is worth paying more attention to? This is called conditional amplification/suppression (with respect to the outside world) as opposed to unconditional excitatory and inhibitory activity in existing DNNs.

The distinctive role of UCF (as a tuner) quite strongly implicates that it is closely related to consciousness (Bachmann and Anthony, 2014; Phillips et al., 2016). Overall, the proposed CMI theory improves our understanding of the mechanisms responsible to produce coherent thoughts, percepts, and actions, which are well-adapted to different situations and long-term goals. The distinction between different contextual fields (LCF and UCF) is certainly a move in the right direction. However, the presented basic mathematical model and results should be taken with care. The CMI neural model needs to be significantly improved by incorporating the observed network behavior, which conditionally amplifies/suppresses the RF-LCF signals with respect to different external environments.

7. FUTURE RESEARCH DIRECTIONS

Future work aims to quantify the suppression and attenuation of multisensory signals in terms of four basic arithmetic operators (addition, subtraction, multiplication and division) and their various forms (Kay et al., 2017). We will analyze how the information in CANN is decomposed into components unique to each other having multiway mutual/shared information. The ongoing and future work also includes studying the application of CMI theory to a range of real-world problems, including: (i) computational modeling of AV processing in Alzheimer's and Parkinson's diseases, and schizophrenia (Phillips et al., 2015, 2016), (ii) natural human-robot interactions, (iii) low-power neuromorphic chips, (iv) brain-computer interface, and (v) neurofinance. Furthermore, several other areas within psychology and neuroscience could potentially benefit from the proposed theory (e.g., Héricé et al., 2016; Karim et al., 2017). Defense Advanced Research Projects Agency (DARPA) recently announced a USD 2 billion campaign to develop the next wave of artificial intelligence (AI) technologies (Szu et al., 2019). Specifically, DARPA seeks contextual reasoning in AI systems to create more collaborative and trusting partnerships between humans and machines. In this context, the proposed theory represents a step change.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Grant No. EP/M026981/1 and deepCI.org.

ACKNOWLEDGMENTS

The author would like to greatly acknowledge Prof. Amir Hussain and Mandar Gogate from the Edinburgh Napier University for their contributions in implementing state-of-the-art lip-reading driven deep learning approach and

contextual AV switching for speech enhancement, which are published previously and cited here for reference. The author would also like to acknowledge Prof. Bruce Graham, Prof. Leslie Smith, Prof. Bill Phillips from the University of Stirling, Areej Riaz from the London Business School, Prof. Newton Howard from Oxford Computational Neuroscience, and Dr. Mino Belle and Prof. Andrew Randall from the University of Exeter for their help and support in several different ways, including appreciation, motivation, and encouragement. Lastly, I would like to greatly acknowledge Prof. Amar Aggoun and the University of Wolverhampton for supporting quality research.

REFERENCES

Abel, A., Marxer, R., Barker, J., Watt, R., Whitmer, B., Derleth, P., et al. (2016). "A data driven approach to audiovisual speech mapping," in *Advances in Brain Inspired Cognitive Systems: 8th International Conference, BICS 2016, Proceedings 8* (Beijing: Springer), 331–342.

Adeel, A., Gogate, M., and Hussain, A. (2019a). Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments. *Inform. Fusion*. doi: 10.1016/j.inffus.2019.08.008

Adeel, A., Gogate, M., Hussain, A., and Whitmer, W. M. (2019b) Lip-reading driven deep learning approach for speech enhancement. *IEEE Trans. Emerg. Top. Comput. Intell.* doi: 10.1109/TETCI.2019.2917039

Bachmann, T., and Anthony, G. H. (2014). It is time to combine the two main traditions in the research on the neural correlates of consciousness: C= L × D. *Front. Psychol.* 5:940. doi: 10.3389/fpsyg.2014.00940

Barker, J., Marxer, R., Vincent, E., and Watanabe, S. (2015). "The third 'chime' speech separation and recognition challenge: dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (Scottsdale, AZ: IEEE), 504–511.

Cooke, M., Barker, J., Cunningham, S., and Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.* 120, 2421–2424. doi: 10.1121/1.2229005

Fox, K., and Daw, N. (1992). A model for the action of nmda conductances in the visual cortex. *Neural Comput.* 4, 59–83. doi: 10.1162/neco.1992.4.1.59

Galletti, C., and Battaglini, P. P. (1989). Gaze-dependent visual neurons in area v3a of monkey prestriate cortex. *J. Neurosci.* 9, 1112–1125.

Gelenbe, E. (1993a). G-networks by triggered customer movement. *J. Appl. Probabil.* 30, 742–748.

Gelenbe, E. (1993b). Learning in the recurrent random neural network. *Neural Comput.* 5, 154–164.

Gonzalez, A. J., Stensrud, B. S., and Barrett, G. (2008). Formalizing context-based reasoning: a modeling paradigm for representing tactical human behavior. *Int. J. Intell. Syst.* 23, 822–847. doi: 10.1002/int.20291

Héricé, C., Khalil, R., Moftah, M., Boraud, T., Guthrie, M., and Garenne, A. (2016). Decision making under uncertainty in a spiking neural network model of the basal ganglia. *J. Integr. Neurosci.* 15, 515–538. doi: 10.1142/S021963521650028X

Karim, A. A., Lützenkirchen, B., Khedr, E., and Khalil, R. (2017). Why is 10 past 10 the default setting for clocks and watches in advertisements? A psychological experiment. *Front. Psychol.* 8:1410. doi: 10.3389/fpsyg.2017.01410

Kay, J., Floreano, D., and Phillips, W. A. (1998). Contextually guided unsupervised learning using local multivariate binary processors. *Neural Netw.* 11, 117–140.

Kay, J., Ince, R., Dering, B., and Phillips, W. (2017). Partial and entropic information decompositions of a neuronal modulatory interaction. *Entropy* 19:560. doi: 10.3390/e19110560

Kay, J., and Phillips, W. A. (1997). Activation functions, computational goals, and learning rules for local processors with contextual guidance. *Neural Comput.* 9, 895–910.

Kay, J. W., and Phillips, W. (2011). Coherent infomax as a computational goal for neural systems. *Bull. Math. Biol.* 73, 344–372. doi: 10.1007/s11538-010-9564-x

Kepecs, A., and Raghavachari, S. (2002). "3 State neurons for contextual processing," in *Advances in Neural Information Processing Systems* (Vancouver, BC), 229–236.

Larkum, M. (2013). A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends Neurosci.* 36, 141–151. doi: 10.1016/j.tins.2012.11.006

Larkum, M., and Phillips, W. (2016). Does arousal enhance apical amplification and disamplification? *Behav. Brain Sci.* 39:e215. doi: 10.1017/S0140525X15001867

Larkum, M. E., Nevian, T., Sandler, M., Polsky, A., and Schiller, J. (2009). Synaptic integration in tuft dendrites of layer5 pyramidal neurons: a new unifying principle. *Science* 325, 756–760. doi: 10.1126/science.1171958

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., and Hinton, G. (2020). Backpropagation and the brain. *Nat. Rev. Neurosci.* 1-12.

Lizier, J. T., Bertschinger, N., Jost, J., and Wibral, M. (2018). Information decomposition of target effects from multi-source interactions: perspectives on previous, current and future work. *Entropy* 20:307. doi: 10.3390/e20040307

Massimini, M., Ferrarelli, F., Huber, R., Esser, S. K., Singh, H., and Tononi, G. (2005). Breakdown of cortical effective connectivity during sleep. *Science* 309, 2228–2232. doi: 10.1126/science.1117256

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748.

Patterson, M. L., and Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Dev. Sci.* 6, 191–196. doi: 10.1111/1467-7687.00271

Phillips, W., Clark, A., and Silverstein, S. M. (2015). On the functions, mechanisms, and malfunctions of intracortical contextual modulation. *Neurosci. Biobehav. Rev.* 52, 1–20. doi: 10.1016/j.neubiorev.2015.02.010

Phillips, W. A. (2001). Contextual modulation and dynamic grouping in perception. *Trends Cogn. Sci.* 5, 95–97. doi: 10.1016/S1364-6613(00)01617-X

Phillips, W. A., Bachmann, T., and Storm, J. F. (2018). Apical function in neocortical pyramidal cells: a common pathway by which general anesthetics can affect mental state. *Front. Neural Circuits* 12:50. doi: 10.3389/fncir.2018.00050

Phillips, W. A., Larkum, M. E., Harley, C. W., and Silverstein, S. M. (2016). The effects of arousal on apical amplification and conscious state. *Neurosci. conscious.* 2016:niw015. doi: 10.1093/nc/niw015

Phillips, W. A., and Silverstein, S. (2013). The coherent organization of mental life depends on mechanisms for context-sensitive gain-control that are impaired in schizophrenia. *Front. Psychol.* 4:307. doi: 10.3389/fpsyg.2013.00307

Phillips, W. A., and Singer, W. (1997). In search of common foundations for cortical computation. *Behav. Brain Sci.* 20, 657–683. doi: 10.1017/S0140525X9700160X

Ross, D. A., Lim, J., Lin, R.-S., and Yang, M.-H. (2008). Incremental learning for robust visual tracking. *Int. J. Comput. Vis.* 77, 125–141. doi: 10.1007/s11263-007-0075-7

Salinas, E., and Sejnowski, T. J. (2001). Correlated neuronal activity and the flow of neural information. *Nat. Rev. Neurosci.* 2:539. doi: 10.1038/35086012

Stein, B. E., and Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nat. Rev. Neurosci.* 9:255. doi: 10.1038/nrn2331

Stein, B. E., Stanford, T. R., and Rowland, B. A. (2009). The neural basis of multisensory integration in the midbrain: its organization and maturation. *Hear. Res.* 258, 4–15. doi: 10.1016/j.heares.2009.03.012

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215.

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica* 36, 314–331.

Szu, H., Chang, L., Chu, H., Kolluru, R., Foo, S. F., and Wu, J. (2019). The 3rd wave AI requirements. *MOJ App. Bio. Biomech.* 3, 18–22. doi: 10.15406/mojabb.2019.03.00094

Viola, P., and Jones, M. (2001). "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001*, Vol. 1 (Kauai, HI: IEEE), I.

Wibral, M., Priesemann, V., Kay, J. W., Lizier, J. T., and Phillips, W. A. (2017). Partial information decomposition as a unified approach to the specification of neural goal functions. *Brain Cogn.* 112, 25–38. doi: 10.1016/j.bandc.2015.09.004

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.