Journal of Neural Engineering

PAPER

Effect of visual input on syllable parsing in a computational model of a neural microcircuit for speech processing

Anirudh Kulkarni[1], **Mikolaj Kegler**[1] and **Tobias Reichenbach**[1,2,*]

[1] Department of Bioengineering and Centre for Neurotechnology, Imperial College London, South Kensington Campus, SW7 2AZ London, United Kingdom
[2] Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Konrad-Zuse-Strasse 3/5, Erlangen, 91056, Germany
* Author to whom any correspondence should be addressed.

E-mail: tobias.j.reichenbach@fau.de

Keywords: input, syllable, parsings, computational, model, neural

OPEN ACCESS

RECEIVED
14 June 2021

REVISED
20 August 2021

ACCEPTED FOR PUBLICATION
21 September 2021

PUBLISHED
12 October 2021

Abstract

Objective. Seeing a person talking can help us understand them, particularly in a noisy environment. However, how the brain integrates the visual information with the auditory signal to enhance speech comprehension remains poorly understood. *Approach.* Here we address this question in a computational model of a cortical microcircuit for speech processing. The model consists of an excitatory and an inhibitory neural population that together create oscillations in the theta frequency range. When stimulated with speech, the theta rhythm becomes entrained to the onsets of syllables, such that the onsets can be inferred from the network activity. We investigate how well the obtained syllable parsing performs when different types of visual stimuli are added. In particular, we consider currents related to the rate of syllables as well as currents related to the mouth-opening area of the talking faces. *Main results.* We find that currents that target the excitatory neuronal population can influence speech comprehension, both boosting it or impeding it, depending on the temporal delay and on whether the currents are excitatory or inhibitory. In contrast, currents that act on the inhibitory neurons do not impact speech comprehension significantly. *Significance.* Our results suggest neural mechanisms for the integration of visual information with the acoustic information in speech and make experimentally-testable predictions.

1. Introduction

Speech comprehension can benefit from other sensory input, in addition to the auditory signal, through multisensory integration [1, 2]. As a striking example, seeing a speaker's face and their moving lips can improve the comprehension of speech in noise by more than 10 dB in the signal-to-noise ratio [3, 4]. Such audiovisual enhancement of speech comprehension may result from different visual features such as facial gestures, hand movements, jaw movements, as well as the alternating configuration of the lips, teeth, tongue, head and eyebrows [5–7]. In particular, the area of the mouth opening is strongly correlated with the amplitude fluctuations in speech, and mouth movements typically precede the corresponding voice onset by about 100–300 ms [8].

Classic theories of such multisensory processing posit that primary sensory regions process only unisensory inputs [9, 10]. The individual streams of information are then relayed to higher-level association cortices where the information from the various unisensory regions converge to create a multisensory percept. However, recent studies in several species have shown that the integration of auditory information with other sensory modalities can occur in the brain as early as the primary and secondary auditory cortices which were hitherto considered to be unisensory areas [11]. For instance, in adult rhesus monkeys, visual stimuli were found to modulate the activity of single neurons as well as the local field potential (LFP) in the primary auditory cortex [12–15]. Similarly, single-unit recordings as well as the LFP in the auditory cortex of anaesthetized ferrets were influenced by visual stimuli [16, 17]. In awake mice, using multisite probes to sample single units across multiple cortical layers, it was demonstrated that visual stimuli influenced

firing in the primary auditory cortex [18] and short-term visual deprivations led to enhanced neuronal responses and frequency selectivity to sounds in layer four of the primary auditory cortex (A1) [19]. Experiments using a voltage-sensitive dye and optical imaging in guinea pigs observed inhibitory responses in auditory areas about 110 ms after the onset of a visual stimuli [20].

In support of multisensory processing in early sensory areas, it has further been demonstrated that direct projections from visual areas to the auditory cortex exist in monkeys [21, 22], ferrets [23], Mongolian gerbils [24], marmosets [25] and they have also been suggested in rats [26].

In humans, using functional magnetic resonance imaging (fMRI), it has similarly been found that visual stimulation and the reading of text by themselves activated the auditory cortex [27, 28]. Further studies using magnetoencephalography showed that viewing a speaker's face improved the tracking of speech rhythms in the auditory cortex [29, 30], and additionally, using intercranial electroencephalography (iEEG), it was shown that the phase of the slow oscillations in the auditory cortex could track the rhythms in a talking face [31]. Moreover, intracranial stereotactic electroencephalographic (sEEG) recordings in human patients suggest direct pathways linking early visual and auditory regions and that visual input is processed in the auditory cortex about 100 ms after the visual onset [32].

Current theories of speech processing include a role of the cortical tracking of the amplitude fluctuations in speech by the different cross-coupled neural oscillations such as delta (1–4 Hz), theta (4–8 Hz) and gamma (25–100 Hz) rhythms [33, 34]. These oscillations occur at the rhythms set by words, syllables and phonemes, respectively. In particular, the theta band is assumed to parse speech into syllables [35–38] thus providing temporal frames for the phonemic encoding by the gamma rhythm. A computational model of a spiking neural network for speech processing that included theta oscillations coupled to gamma oscillations showed that phonemes could indeed be decoded from the gamma activity when it was parsed by the input from the theta oscillator.

Visual enhancement of speech comprehension may, at least in part, result from the visual stimuli affecting oscillatory activity in the auditory cortex. Studies on ferrets showed that information from the visual cortex was conveyed to the auditory cortex by influencing the phase of the LFP [17]. One study found increases in alpha power in the auditory cortices due to visual signals [39] whereas others observed changes, including phase resets, in the delta (3–4 Hz), theta (4–8 Hz), beta (12–30 Hz) and alpha (8–14 Hz) frequency bands [32, 33, 40–42].

The precise mechanisms by which visual signals can influence cortical oscillations related to speech, and thereby impact speech comprehension remain, however, elusive. Even though there have been computational models of phase resets of delta oscillations [43] and biophysical models of phase locking of oscillators [44, 45], none of them investigated how these effects relate to speech processing.

In this study we employ a recently suggested model of a spiking neural network for speech processing to investigate the effect of visual input [46]. In particular, the artificial neural network includes a module for theta oscillations that can parse speech into distinct syllables. We investigate how different types of speech-related visual input influence the accuracy of the syllable parsing.

2. Methods

2.1. Architecture of the computational model

Our artificial spiking neural network for speech processing is based on a recently-introduced model that contains coupled theta- and gamma-oscillations [46]. The theta oscillations thereby segment a speech stream into individual syllables, and the neural activity in the gamma range allows us to decode the syllable identity.

The auditory speech input is first processed by a model of the thalamus before reaching a module that produces oscillations in the theta range (figure 1(a)). Because we are interested in investigating the influence of slow visual input, such as related to the opening and closing of the mouth, on speech processing, our model includes only the theta oscillator and not also a gamma oscillator. When stimulated by a speech input, the spiking activity of the theta module becomes aligned to the syllable boundaries. An example speech input, its time frequency spectrogram and the resulting LFP and spiking neural activity are shown in figure 1(b).

The theta module produces oscillations through an interplay of excitatory neurons (Te) and inhibitory neurons (Ti) that are reciprocally coupled via inhibitory and excitatory synapses. The theta-band oscillations are generated by the principle of slower feedback inhibition following fast recurrent excitations. At the beginning of each oscillatory cycle, the excitatory input increases, resulting in an increase in the firing rate of the excitatory population. The inhibitory population eventually catches up and brings down the firing rate of the excitatory population. As the excitatory population activity goes down, and as a result the inhibitory population activity decreases, the network recovers from inhibition and the excitatory firing rate increases again. This results in a rhythmic behaviour that is referred to as the pyramidal interneuron theta (PIN-TH) mechanism, analogous with the pyramidal interneuron gamma model [47].

We consider ten excitatory neurons that are reciprocally connected to each other. Likewise, we model

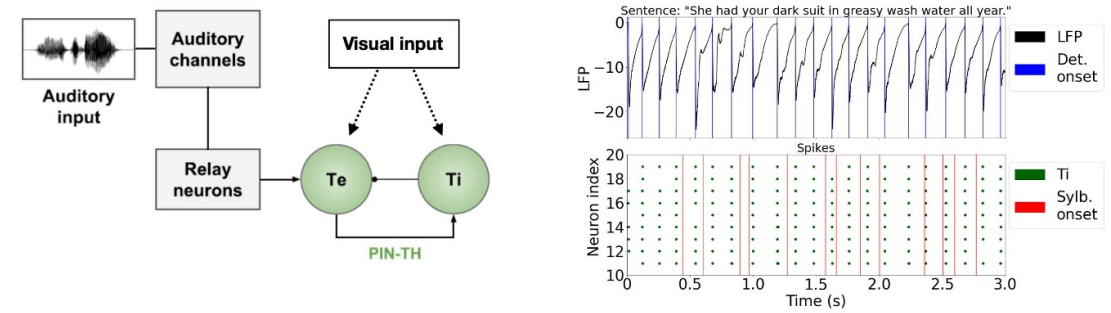

Figure 1. Architecture of the spiking neural network and the extraction of syllable onsets. (a) Network architecture. The auditory input is decomposed through 32 frequency channels and the resulting signal is relayed through a population of relay neurons, which act as a spectro-temporal filter, to the theta module. The theta module consists of ten excitatory neurons (Te) and ten inhibitory neurons (Ti) and generates self-sustained oscillations in the theta frequency band. The visual input is added to either Te or Ti. (b) The theta LFP generated by an example sentence 'She had your dark suit in greasy wash water all year' together with the estimated syllable onsets (blue, top panel). The spiking of the inhibitory Ti neurons (blue dots) is aligned to the syllable onsets (red lines, bottom panel).

ten inhibitory neurons with all-to-all connections as well. The all-to-all connectivity within the Te neurons, respectively, the Ti neurons, means that we model a local cortical network.

The neurons are modelled as leaky integrate-and-fire neurons with the following dynamics for the voltage V_i for cell i:

$$C\frac{\mathrm{d}V_i}{\mathrm{d}t} = g_{\mathrm{L}}\left(V_{\mathrm{L}} - V_i\right) + I_i^{\mathrm{SYN}}(t) + I_i^{\mathrm{Inp,aud}}(t) + I_i^{\mathrm{Inp,vis}}(t) + I_i^{\mathrm{DC}} + \eta(t) \quad (1)$$

where C is the capacitance of the cellular membrane; g_{L} and V_{L} are the conductance and the reversal potential of the leak current; $I_i^{\mathrm{SYN}}(t)$, $I_i^{\mathrm{Inp,aud}}(t)$, $I_i^{\mathrm{Inp,vis}}(t)$, I_i^{DC} are the synaptic current, the auditory stimulus-induced current, the visual stimulus-induced current and the constant direct current delivered to the cell. $\eta(t)$ is white Gaussian noise with a variance of σ^2. Whenever the membrane potential of the neuron reaches the threshold potential V_{THR}, a spike is generated and returned to the reset potential V_{RESET}.

The synaptic current $I_{ij}^{\mathrm{SYN}}(t)$ to the postsynaptic neuron i from the presynaptic neuron j is modelled as follows:

$$I_{ij}^{\mathrm{SYN}}(t) = g_{ij}s_{ij}(t)\left(V_j^{\mathrm{SYN}} - V_i(t)\right) \quad (2)$$

where g_{ij} is the conductance of the synapse connecting neuron j to neuron i; $s_{ij}(t)$ is the activation variable of the synapse, and V_j^{SYN} is the equilibrium potential of the synaptic current from neuron j.

The dynamics of the activation variables $s_{ij}(t)$ of the neurons are described by the following set of equations:

$$\frac{\mathrm{d}x_{ij}^{\mathrm{R}}}{\mathrm{d}t} = -\frac{x_{ij}^{\mathrm{R}}}{\tau_j^{\mathrm{R}}} + \delta\left(t - t_j^{\mathrm{SPK}}\right) \quad (3)$$

$$\frac{\mathrm{d}s_{ij}}{\mathrm{d}t} = \frac{x_{ij}^{\mathrm{R}} - s_{ij}}{\tau_j^{\mathrm{D}}} \quad (4)$$

where x_{ij}^{R} are activation variables of the synapse from neuron j to neuron i; $\delta\left(t - t_j^{\mathrm{SPK}}\right)$ denotes a spike generation in a presynaptic neuron j at the time t_j^{SPK}, and τ_j^{R} and τ_j^{D} are time constants of synaptic rise and decay of the presynaptic neuron j, respectively.

Therefore, $I_i^{\mathrm{SYN}}(t)$, the sum of all synaptic inputs from the cells projecting to the ith neuron, is given by:

$$I_i^{\mathrm{SYN}}(t) = \sum_j g_{ij}s_{ij}(t)\left(V_j^{\mathrm{SYN}} - V_i(t)\right). \quad (5)$$

The LFP at time t, LFP(t), is obtained by summing the absolute values of all the synaptic currents delivered to all the theta excitatory cells in the network [48].

The model parameters were adapted from [49]. The complete list of model parameters and their values are presented in table 1. All numerical simulations of the model were performed in a custom written Python script using the packages SciPy [50] and Brian2, a Python package for implementing simulations of networks of neurons [51]. We used a time step of 0.01 ms in all our simulations.

2.2. Auditory stimuli and their processing in the model

Spoken English sentences from either the TIMIT dataset [52] or from the GRID corpus [53] were provided as the auditory input to the network. The TIMIT corpus reflects realistic listening scenarios by incorporating speakers of different accents and speech production rates. It comprises of over 6300 phonetically-labelled sentences. The GRID corpus, on the other hand, contains both the audio and visual recordings of 34 speakers speaking 1000 sentences each.

The material from the TIMIT corpus was used for the simulations with a pulse input current, whereas

Table 1. Model parameters.

Parameter	Description	Value
	Neuron model	
C	Cell membrane capacitance	1 pF
V_{THR}	Spiking threshold	−40 mV
V_{RESET}	Resting potential	−87 mV
V_E^{SYN}	Equilibrium potential of excitatory neurons	0 mV
V_I^{SYN}	Equilibrium potential of inhibitory neurons	−80 mV
	PIN-TH network	
g_{LE}	Leak conductance in Te neurons	0.0264 nS
g_{LI}	Leak conductance in Ti neurons	0.1 nS
τ_{Te}^R	Synaptic rise constant of Te neurons	4 ms
τ_{Ti}^R	Synaptic rise constant of Ti neurons	5 ms
τ_{Te}^D	Synaptic decay constant of Te neurons	24.3 ms
τ_{Ti}^D	Synaptic decay constant of Ti neurons	30.36 ms
I_{Te}^{DC}	Constant current delivered to Te neurons	1.25 pA
I_{Ti}^{DC}	Constant current delivered to Ti neurons	0.0851 pA
σ_{Te}	Variance of the noise term in Te neurons	0.282 pA ms$^{1/2}$
σ_{Ti}	Variance of the noise term in Ti neurons	2.028 pA ms$^{1/2}$
	Connectivity	
$g_{Te,Ti}$	Ti → Te synaptic conductance strength	$2.07/N_{Ti}$ nS
$g_{Ti,Te}$	Te → Ti synaptic conductance strength	$3.33/N_{Te}$ nS
$g_{Ti,Ti}$	Ti → Ti synaptic conductance strength	$4.32/N_{Ti}$ nS

the data from the GRID corpus and the corresponding videos were used for simulations where the visual current corresponded to either the area of the mouth opening or its velocity. A silent period sampled from a uniform distribution in the range 250–750 ms was added to each sentence to provide variability in the onset of the sentence with respect to the intrinsic firing of the theta module. This was done in order to avoid any spurious phase-locking of the network rhythm to the speech input rate.

For each simulation, a random subset of 100 sentences were chosen as the speech input in the model simulation. Speech-shaped noise was then added to each of these speech inputs. To produce the speech shaped noise, another randomly selected sentence was picked from the TIMIT database. From the linear prediction coefficients of this second sentence, a linear filter was computed. The linear filter was then convolved with a white-noise Gaussian signal to yield the speech shaped noise. The speech signal and the resulting speech shaped noise signal were mixed at a signal-to-noise ratio (SNR) of 0 dB to produce the auditory input to the model.

The auditory input was first processed through the model of peripheral and subcortical auditory processing [54]. The subcortical model represented the cochlear filter bank and decomposed the input signal into 128 auditory channels with centre frequencies that are logarithmically spaced between 100 and 4000 Hz [54]. A series of non-linear operations representing the neural processing in the auditory nerve and subcortical nuclei were then performed on this decomposed signal. The model was implemented in a custom written Python script based on the original MATLAB implementation [46].

The number of auditory channels was then reduced to 32 by taking every fourth channel from the 128 channels. In order to reflect the experimental observation of the entraining of endogenous theta activity in the auditory cortex to the syllabic rhythm of natural speech stimuli, the theta module was designed to generate bursts of spikes aligned to the syllabic onsets in the presented sentence. For this purpose, the 32 obtained auditory channels were convolved with a spectro-temporal filter and projected to the Te neurons. This spectro-temporal filter represented a population of relay neurons with weights that corresponded to the synaptic strengths [55]. It projected the inputs with a delay of up to 50 ms and predicted syllabic onsets (binary events) based on the data from the 32 auditory channels from up to 50 ms preceding time t, in steps of every 10 ms:

$$\hat{Y}(t) = \sum_{c=1}^{32} \sum_{\tau=-50}^{0} B(c,\tau) X(c, t+\tau). \qquad (6)$$

$Y(t)$ is a binary variable indicating the syllabic onsets in a sentence; $\hat{Y}(t)$ is an estimate of that variable; c is the index of the auditory channel; τ is the latency in ms with respect to time t; B is a matrix of filter coefficients and X is the input from auditory channel c at time t. The binary vector $Y(t)$ was determined such that it had a value of one at the onset of each syllable but was zero elsewhere.

To obtain the coefficients B of this spectro-temporal filter, 1000 sentences that were not subsequently used for any simulations of the network, were randomly chosen from the TIMIT corpus. These sentences were appended with a silence of 500–1000 ms at the beginning and were processed through

the above-described auditory periphery model and then downsampled to 100 Hz and concatenated to give X. The binary vector with the corresponding syllabic onsets were processed accordingly to obtain Y. The coefficients B were then obtained by providing an optimal mapping between X and Y using sparse bilinear regression [56]. Once the filter coefficients were obtained, we convolved the optimized kernel with the 32 auditory channels and scaled it down, to regulate the input current, by a factor of 4.5 to obtain the input to the Te neurons as in the original model by [46].

2.3. Syllable parsing in the model

The speech input was added in the model through a current, $I_i^{\mathrm{Inp,aud}}$, to the excitatory neurons (Te), as specified above. The visual input, on the other hand, was added to the model through the visual current term $I_i^{\mathrm{Inp,vis}}$ either to the pyramidal neurons (Te) or to the inhibitory neurons (Ti).

In the absence of any speech input, the model exhibited self-sustained theta oscillations. When auditory input was added, this signal was chunked into distinct units by the theta rhythm. In particular, these chunks were delineated by the rhythmic spike bursts in the theta inhibitory module and were considered to represent individual syllables. A theta spike burst was thereby considered to be represented by the spiking of at least two inhibitory neurons in the theta module within a time window of 15 ms. The timing of such a spike burst was then considered to be the time of the maximal firing rate of Ti neurons.

2.4. Analysis of syllable parsing in the model

To quantify the accuracy of the model's syllable parsing, we computed a distance measure between the syllable boundaries inferred from the network activity and the actual boundaries, called the parsing score. The parsing score was obtained in three steps: (a) we computed a distance metric between the model's predictions and actual syllabic onsets, (b) we subtracted a control distance from this measure and (c) we divided the net result by the number of syllables. A parsing score of 1 therefore corresponded to perfect parsing by the model and a parsing score of 0 is what one would expect by chance.

To compute the distance metric in the first step, we used the normalized Victor–Purpura spike distance metric (VPd) [57] to quantify the overall misalignment of the predicted and the actual syllable onsets. Misalignment can result from missed syllable onsets, misaligned onsets, or additional onsets inferred from the network activity. The VPd is particularly suitable for this task (and commonly used in spike train analysis) because it captures all three types of misalignments. The VPd between two series of binary events is calculated as the minimum cost of transforming one series into the other using one of the three operations: insertion of an event, deletion of an event and shifting of an event. A cost parameter of 50 ms was used. Hence, when the timing difference between the predicted and the actual syllable boundaries was no more than 50 ms, the two were said to be matched. A value corresponding to the ratio of the time difference to the cost was added to the distance parameter. When they were more than 50 ms apart, the score was augmented by 1. This was then subtracted from a control score defined by the normalized VPd score between the syllabic onsets and uniformly distributed bursts of spikes in the same interval. The onset in the case of the control score calculation was chosen in the same way as the random onset of the sentence. The difference in the earlier distance score and the control score was then divided by the number of syllables of the sentences, to normalize the score across different acoustic speech inputs.

The parsing scores were obtained in the same way for every simulation irrespective of the external audio and visual current inputs. The analysis was implemented in a custom written Python script using methods from SciPy package. The significance of the parsing scores obtained in each visual input condition with respect to the no-visual condition was computed using the Wilcoxon signed-rank test [58]. We then applied the Benjamini–Hochberg correction to the obtained p-values to check for false discoveries from multiple comparisons [59]. The significance threshold for the hypothesis testing was set to $p = 0.05$.

2.5. Extraction of mouth area from the videos

To extract visual information from the videos of the GRID corpus, and in particular the mouth area, we used a custom-written Python script. The videos of the GRID corpus typically had the face of a speaker on a blue background while the speaker recited the sentence. The videos had a frame rate of 25 Hz and the speakers' face had been aligned across all the frames of the video. Each image was cropped to a small region around the mouth. The corresponding cropping region was manually determined for each speaker and stayed the same throughout the video. The pixels of the lips were then extracted using the property that the intensity of the red hue of these pixels was generally greater than the intensities of the blue or green pixels. The image was then blurred with a Gaussian filter to remove small, isolated pixels. By extracting connected objects greater than a certain threshold, we could thus extract the outer boundary of the lips. An example of the extracted lip contour whose outer boundary is highlighted by green dots is shown in figure 3(a). The number of pixels enclosed within this outer boundary was then computed for each image to obtain the area of the open mouth. We z-scored the number of pixels and upsampled the resulting signal to the same frequency as the auditory signal used in the simulation of the model corresponding to the time step of the simulation of 0.01 ms, i.e. 100 KHz, to obtain the mouth-opening area. An

example is shown in figure 3(b). To explore the influence of the magnitude of the visual input on the audiovisual speech processing, in certain simulations, we multiplied the resulting signal with a factor that we called the amplitude of the area of the mouth opening. Effectively, this corresponded to scaling the standard deviation of the mouth area signal. An amplitude of one was used for the current unless mentioned otherwise.

2.6. Extraction of velocity of the mouth-opening area from the videos

To obtain the velocity of the mouth-opening area, we computed the time difference of the number of pixels within the lip contour that we obtained for a given speaker in a video. This difference signal was then z-scored to obtain the velocity of the mouth-opening area. In certain simulations, we multiplied this resulting signal with a certain factor, that we termed the amplitude of the velocity of the mouth-opening area. Effectively, this corresponded to scaling the standard deviation of the signal. An amplitude of one was used for the current unless otherwise mentioned. An example signal of the velocity of the mouth area obtained is shown in figure 4(a).

2.7. Adding visual input to this network

We incorporated a visual input current of one of the three types. First, we considered a rectangular pulse current in which each pulse had a duration of 25 ms and was located at the syllable boundary. This current represented a simplified input that bore direct relation to the syllable rhythms in the speech signal. Second, we considered a current that varied in proportion to the mouth-opening area. This current represented an important feature of the visual stimuli. Third, we investigated a current that was proportional to the velocity of the mouth-opening area. This current was chosen since the visual cortex can extract motion aspects from videos.

We did not consider further, more complex spatiotemporal filters for the video signal. Unlike the spectro-temporal auditory filter, such visual filters would most likely perform poorly, due to the much higher dimensionality of the visual input. As the flow of visual information to the speech processing areas of the brain remains poorly understood, we have chosen this simplified model of lip detection rather than a more intricate representation of the visual signal.

These visual currents were added as $I_i^{\mathrm{Inp,vis}}(t)$ to either the excitatory population or the inhibitory population of the theta network module. Moreover, the visual input current was offset in time with respect to the corresponding auditory signal such that we could investigate the effects of the different time-lagged offsets in visual current on the syllable parsing scores.

For each of the three different types of currents, we studied four conditions: (a) adding an excitatory visual input current to the excitatory neurons of the theta module, (b) adding an excitatory visual input current to the inhibitory neurons, (c) adding an inhibitory visual input current to the excitatory neurons of the theta module, and (d) adding an inhibitory visual input current to the inhibitory neurons. In all cases, we compared the resulting syllable parsing scores to the condition with no visual input current.

2.8. Computing the phase of the signal

To compute the instantaneous phase of the LFP, we used the Hilbert transform. The LFP from the theta module was firstly filtered using a third-order lowpass Butterworth filter with a cutoff frequency of 30 Hz. The Hilbert transform was subsequently applied to the resulting signal to obtain the envelope and phase of the signal. We then determined the phase of the signal at the syllabic onsets of the sentence. The mean phase at the syllabic onsets was computed under the different visual input conditions and was compared to the case with no visual input.

2.9. Computing the scalogram of the signal

To compute the scalogram of the LFP, we consider the frequency band between 1 and 100 Hz and performed a Morlet continuous wavelet transform with logarithmically spaced frequencies over this frequency band. This function was implemented using the Time Frequency Misfit module in the Signal module of the Obspy package in Python [60]. Once the time-frequency coefficients of the scalogram were obtained, we squared their absolute values and averaged them over time to obtain different coefficients of the average squared scalogram as a function of frequency. This quantity represents the power spectral density in the LFP [61].

3. Results

We first verified that the theta module yielded oscillations in the theta frequency range. We found that, before the beginning of a sentence, the module produced bursts at an interval of about 150 ms (figure 1(b)). These regular bursts of neuronal spikes were also visible in the LFP.

When a speech stimulus was presented to the network, the spiking activity of the theta network became aligned to the syllable onsets (figure 1(b)). This allowed us to investigate how well this syllable parsing through the spikes of the theta module performed, and how this performance changed with the addition of different visual stimuli.

3.1. Syllable parsing score for pulsed input current

We first considered a current that consisted of pulses of a duration of 25 ms and an amplitude of 10 pA, unless mentioned otherwise. Each pulse occurred at the onset of a syllable, although we also considered

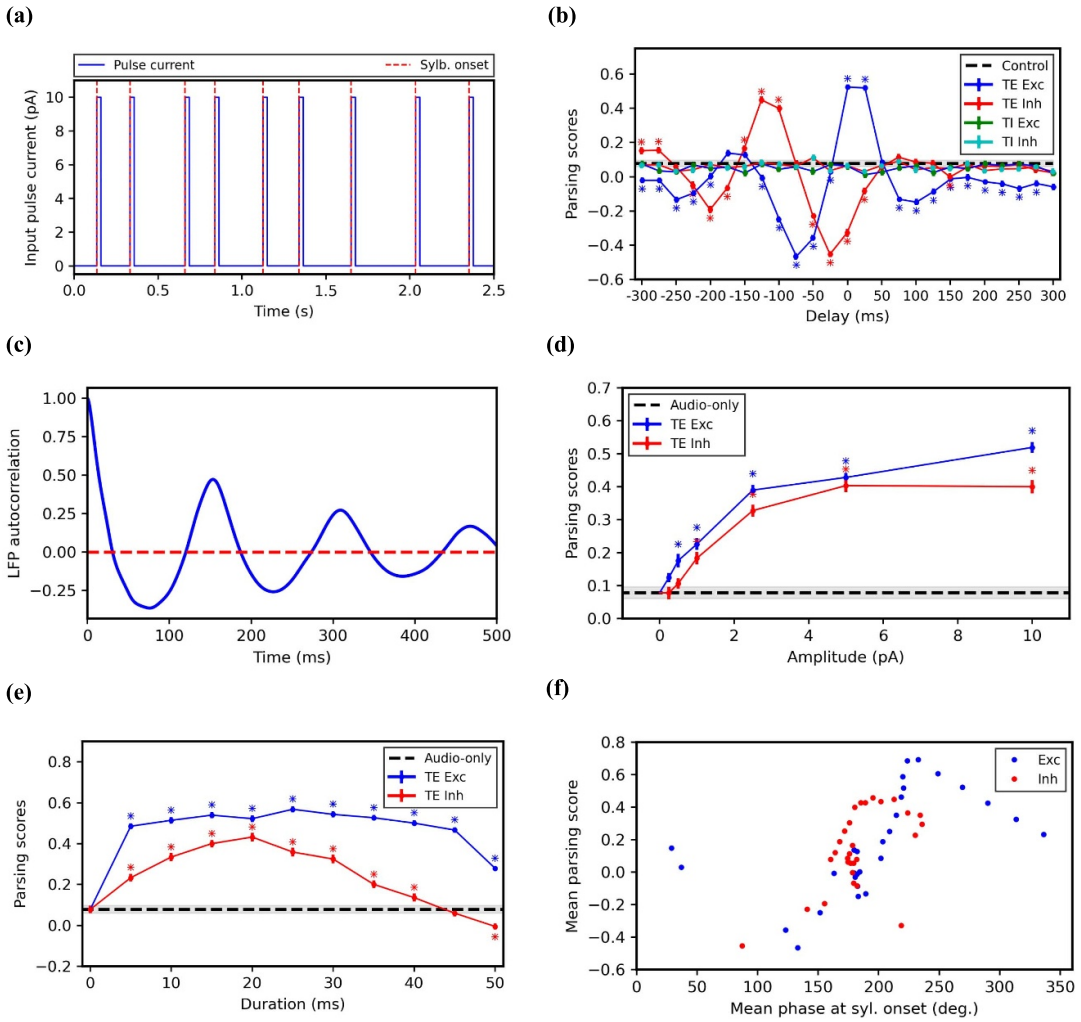

Figure 2. Effect of a pulse input current on the parsing scores. (a) An example of an excitatory pulse input current signal with the onsets of the pulses located at the syllable onsets. (b) An excitatory input to the excitatory neurons (blue) can both improve the syllable parsing or impede it, depending on the delay. A positive delay hereby means that the pulses occur after the corresponding syllable onset. An inhibitory current to the excitatory neurons can influence the syllable parsing as well (red). Neither excitatory (green) nor inhibitory (cyan) current projected to the inhibitory neurons, however, has a significant effect on the syllable parsing. (c) The autocorrelation of the LFP in the absence of a speech stimulus or a visual current shows a periodicity of about 150 ms. (d) The mean parsing scores as a function of the amplitude of the pulse current. The excitatory inputs (blue) occur at 25 ms after the syllable onsets whereas the inhibitory input (red) is presented 125 ms before the syllable onset. (e) The mean parsing scores as a function of the duration of the pulse current. The excitatory inputs (blue) are presented 25 ms after the syllable onsets whereas the inhibitory input (red) occurs 125 ms before the syllable onset. (f) The mean parsing score is optimal for a particular phase of the LFP at the syllable onset, both for the excitatory current (blue) and for the inhibitory current (red). Statistical significance is denoted by asterisks ($p < 0.05$, FDR correction for multiple comparisons).

different time lags between the pulses and the corresponding syllables. An example of an excitatory pulse current where the onsets of the pulses coincided with the syllable onsets is shown in figure 2(a).

The parsing score for speech without any background auditory noise was 0.08, similar to the score obtained in the original model by Hyafil *et al* [46]. In the audio-only condition in our simulations, the audio is comprised of a speech input with background speech-shaped noise at an SNR of 0 dB. This resulted in a parsing score of 0.06 for the audio-only condition. These comparatively low parsing scores reflected frequent missed syllable onsets, misaligned onsets, and additional onsets inferred from the neural activity.

Adding an excitatory pulse current to the excitatory neurons significantly changed the parsing score (figure 2(b)). When the onset of the pulsed coincided with the syllabic onsets, the parsing score improved significantly compared to the audio-only score of 0.06, that is, compared to the condition without any visual input current. On the other hand, delays of around −75 ms as well as around 100 ms led to significantly worse syllable parsing. A positive delay hereby meant that the current pulses occurred after the corresponding syllabic onsets.

When an inhibitory pulse input was presented to the excitatory population, we observed a significant improvement in the mean parsing score at a delay of about −125 ms, as well as a significant

worsening of syllable parsing at a delay of about −25 ms (figure 2(b)).

Adding an excitatory or inhibitory pulse current to the inhibitory neurons of the theta module, however, had no significant effect on the syllable parsing (figure 2(b)). This could have resulted from the recurrent inhibitory connections in the inhibitory population of the network, that may have effectively stunted the activity of the neurons in spite of an external input current.

The parsing scores showed a periodicity of about 150 ms as a function of the delay of the input pulse current when the latter was presented to the excitatory neurons. This periodicity was comparable to the periodicity in the LFP of the theta module, as evident in the autocorrelation of the LFP without a visual or speech input (figure 2(c)). Adding an excitatory pulse current at the syllabic onset presumably made the neurons ready to fire at the syllabic onset. An inhibitory current, on the other hand, reset the excitatory population, such that the neurons were ready to fire together in the next theta cycle, at the syllabic onset.

Next, we investigated the effect of the amplitude and duration of the input current to the excitatory neurons on the parsing scores. To this end, we considered two time lags in the input current: a delay of 25 ms in the case of the excitatory input, and an advance of 125 ms for the inhibitory pulses. These time lags were chosen because they produced the largest significant improvements in the parsing score for the respective currents. The parsing scores improved with the amplitude of the visual current, in particular for smaller currents below 3 pA (figure 2(d)).

To vary the duration of the pulses, we fixed the onset of the pulse current at a delay of 25 ms for the excitatory input, and at an advance of 125 ms for the inhibitory current. We then varied the location of the offset, thus varying the duration of the pulse. The mean parsing score improved as the duration of the pulse increased until 25 ms and then reduced again for longer durations (figure 2(e)). The pulse current presumably reset the activity of the population, and the reset may have been more efficient for longer pulses. However, the longer each pulse lasted, i.e. the further the offset of the pulse current was, the more delayed the reset of the theta population likely was, thus delaying the matching of the theta prediction with respect to the actual syllabic onset. This effect may have caused the degradation of the parsing score for longer pulses.

To explicitly test this hypothesis, we computed the mean phase of the LFP at the syllabic onset for the different pulse current stimulations. We then related the phase of the LFP to the parsing score (figure 2(f)). The parsing scores showed a strong dependency on the phase. In particular, the parsing score improved most when the phase of the oscillation was reset to about 270° for the excitatory input, and to about 200° for the inhibitory current.

3.2. Mouth-opening area

Next, we added a visual input current that corresponded to the area of the mouth (figures 3(a) and (b)). We thereby considered both positive and negative amplitudes. A positive amplitude hereby meant an excitatory current, and a negative amplitude an inhibitory current. We also considered different delays between the mouth-opening current and the speech signal.

Presenting both an excitatory or an inhibitory current corresponding to the mouth-opening area to the theta excitatory population could increase as well as decrease the parsing score, depending on the delay (figure 3(c)).

In particular, an excitatory current led to a worsening of syllable parsing at delays of around 50 ms. An inhibitory current at a delay of about 50 ms led to enhanced syllable parsing, whereas delays of around −150 and 150 ms led to lower parsing scores.

The observed temporal dependencies resembled the ones obtained in a computational model on neurostimulation with the speech envelope [49]. This similarity may result from the considerable correlation of the mouth-opening area and the speech envelope of 0.4 [8].

In contrast, adding such a current to the theta inhibitory population did not affect the parsing score.

We also investigated how the improvements in the parsing score for the current presented to the excitatory neurons varied with the amplitude of the current (figure 3(d)). We thereby considered a delay of −125 ms for the excitatory current, and a delay of 50 ms for the inhibitory one. The excitatory current only produced a significant change in the parsing score at the highest amplitude. In contrast, the inhibitory current showed significantly improved syllable parsing only for small amplitudes. The latter effect may imply that this current must be of the same order as that of the speech signal in order to improve syllable parsing.

3.3. Velocity of the mouth opening area

The third set of visual stimuli that we considered corresponded to the velocity of the mouth-opening area (figure 4(a)). We multiplied this current with a certain value that we refer to as the amplitude of the current. A positive amplitude results in an excitatory current, and a negative amplitude in an inhibitory one. This current was also offset in time by different delays with respect to the corresponding auditory speech input.

As for the two other types of current, we found that both excitatory and inhibitory currents presented to the excitatory neurons could increase as well as decrease the parsing scores (figure 4(b)). In particular, an excitatory current at a delay of about −50 ms increased the parsing score, whereas a delay of about 100 ms led to a decrease. An inhibitory current could

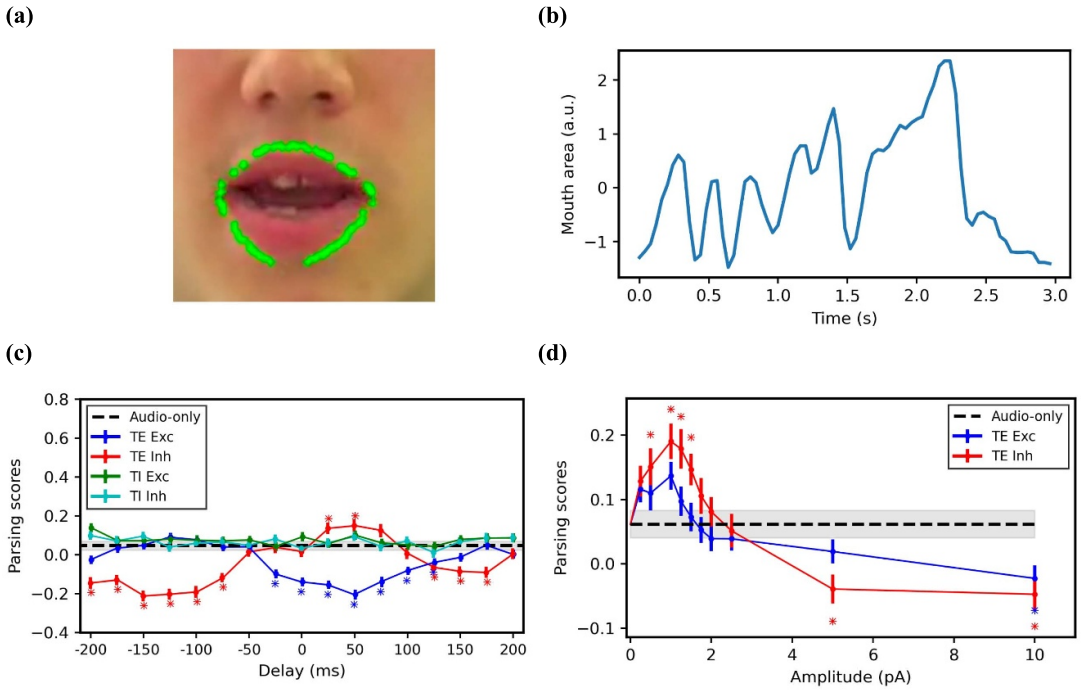

Figure 3. Effect of a current proportional to the mouth-opening area on the parsing scores. (a) The area of mouth opening is derived from the contour of the lips (green) (b) The area is then computed for every image in a video and z-scored to yield the time-varying mouth-opening area. (c) Adding an excitatory (blue) current to the excitatory neurons could significantly worsen the parsing score, whereas an inhibitory current (red) could both improve and worsen it. In contrast, neither an excitatory nor an inhibitory current presented to the inhibitory neurons had an impact on the paring scores (green and cyan). (d) The mean parsing scores as a function of the amplitude of the visual current. The excitatory input (blue) is presented 125 ms preceding the auditory input whereas the inhibitory input (red) is added at a delay of 50 ms. Statistical significance is denoted by asterisks ($p < 0.05$, FDR correction for multiple comparisons).

enhance syllable parsing at a delay of 125 ms and worsen the syllable parsing when presented at a delay of about −50 ms.

When presented to the inhibitory neurons, however, such currents had no significant effect on syllable parsing (figure 4(b)).

We also explored how the amplitudes of the currents, presented to the excitatory neurons, influenced the parsing scores (figure 4(c)). We thereby considered a delay of −50 ms for the excitatory current, and a delay of 125 ms for the inhibitory current. As we observed for the current that was based on the mouth-opening area, large amplitudes of the current degraded the parsing score.

3.4. Firing rates of the excitatory neurons under different visual input conditions

As detailed above, we found that different types of visual currents can enhance syllable parsing when presented to the excitatory neurons. In particular, these current were (a) an excitatory pulse current with a delay of 25 ms, (b) an inhibitory pulse current at a delay 125 ms, (c) an excitatory current corresponding to the mouth-opening area at a delay of −125 ms, (d) an inhibitory current corresponding to the mouth-opening area at a delay of 50 ms, (e) an excitatory current corresponding to the velocity of the mouth-opening area at a delay of −25 ms, and (f) an excitatory current corresponding to the velocity of the mouth-opening area at a delay of 125 ms.

We wondered how the firing rates of the excitatory neurons changed during the presentation of these currents (figure 4(d)). We found that all currents produced changes in the firing rates as compared to the lack of a visual current. Most currents led to moderately higher firing rates. However, the excitatory pulse current caused a much larger firing rate, more than twice the one obtained without visual input. The inhibitory pulse current, on the other hand, yielded a somewhat lower firing rate than without visual current.

3.5. Spectrogram and scalogram data for the different conditions

As another assessment of the effects of the visual currents on the network activity, we investigated the LFP as well. We computed the squared absolute values of the spectrogram and determined how they varied as a function of frequency for the different input visual currents (figure 5). This quantity is the analogue of the power spectral density for the wavelet transform. An example LFP, the corresponding time-frequency spectrogram and the average of the squared absolute value of the spectrogram in the case of no-visual input is shown in figure 5(a).

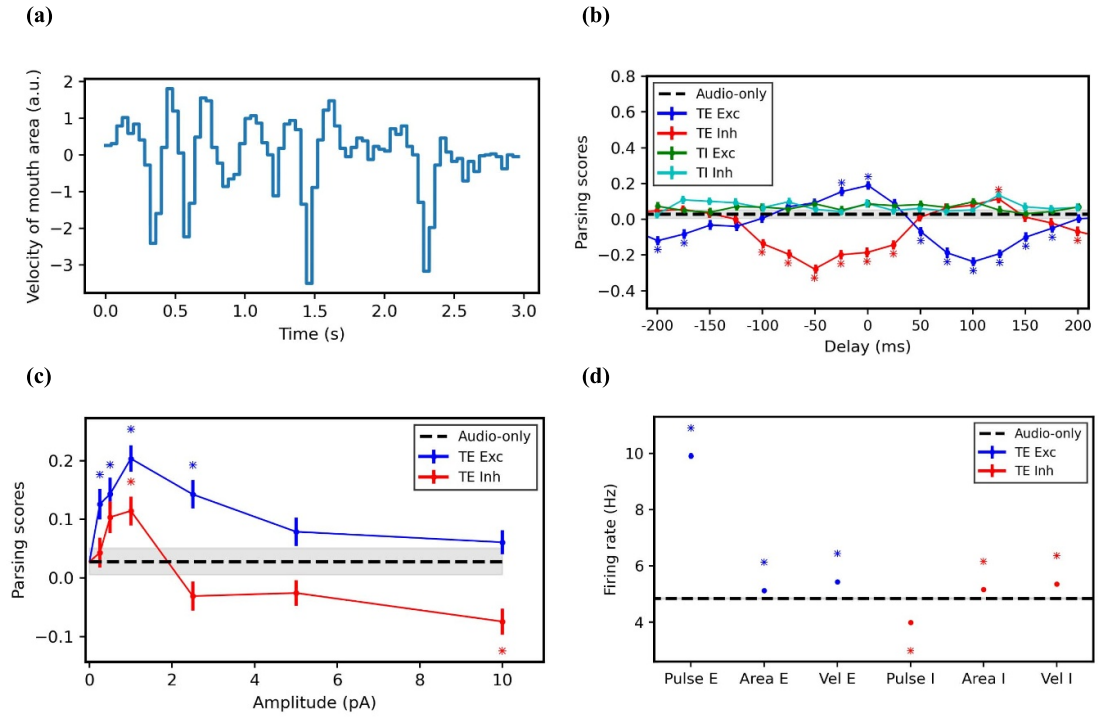

Figure 4. Effect of a current based on the velocity of the mouth-opening area on the parsing scores. (a) An example current signal. (b) Presenting an excitatory current (blue) or an inhibitory current (red) to the excitatory neurons could, at certain delays, significantly improve the parsing score as compared to no visual input (black). However, adding either an excitatory or inhibitory current to the inhibitory neurons (cyan and green) did not influence the syllable parsing. (c) The mean parsing scores as a function of the amplitude of the visual current. The excitatory inputs (blue) were presented at a delay of −25 ms, whereas the inhibitory input (red) was added at a delay of 125 ms. (d) The firing rates of the excitatory neurons under the different conditions. The excitatory pulse inputs (Pulse E) were presented at a delay of 25 ms, whereas the inhibitory pulse inputs (Pulse I) occurred at a delay 125 ms. The excitatory current corresponding to the mouth-opening area (Area E) was presented at a delay of −125 ms whereas the corresponding inhibitory current input (Area I) occurred at a delay of 50 ms. The excitatory current corresponding to the velocity of the mouth-opening area (Vel E) was added at a delay of −25 ms whereas the corresponding inhibitory current (Vel I) occurred at a delay of 125 ms. All input currents cause significantly different firing rates in the excitatory neurons, and in particular for the case of the excitatory pulse inputs (Pulse E). Statistical significance is denoted by asterisks ($p < 0.05$, FDR correction for multiple comparisons).

We then investigated the impact of the six different currents described above that enhanced syllable parsing: (a) the excitatory pulse current with a delay of 25 ms, (b) the inhibitory pulse current at a delay 125 ms, (c) the excitatory current corresponding to the mouth-opening area at a delay of −125 ms, (d) the inhibitory current corresponding to the mouth-opening area at a delay of 50 ms, (e) the excitatory current corresponding to the velocity of the mouth-opening area at a delay of −25 ms, and (f) the excitatory current corresponding to the velocity of the mouth-opening area at a delay of 125 ms.

We found that excitatory pulses increased the overall power of the signal and shifted the location of the maximum from around 6 to 5 Hz while adding a second local maximum at around 12 Hz (figure 5(b)). The inhibitory pulse current also increased the overall power of the signal, though to a smaller extent than the excitatory current, and shifted the maximum slightly to a lower frequency.

The excitatory and inhibitory currents that were based on the mouth-opening area both redistributed the power of the LFP and shifted the location of the maximum to a slightly higher frequency while causing an additional larger peak at a lower frequency (figure 5(c)). The amplitude of the maximum at the lower frequency was slightly higher for the case of the excitatory current than that of the inhibitory current.

Regarding the currents based on of the velocity of the mouth-opening area, both excitation and inhibition increased the power of the LFP and caused an additional maximum at a low frequency (figure 5(d)).

4. Discussion

We studied the effects of visual input on syllable parsing in an artificial neural network for speech processing. The neural network contained a theta module that consisted of coupled excitatory as well as inhibitory neurons and produced rhythmic bursts of spikes in the theta frequency range. When stimulated by speech, the spike bursts became aligned to the syllable boundaries, parsing the speech stream into distinct functional units.

We designed the computational model to explore possible mechanisms of audio-visual integration in

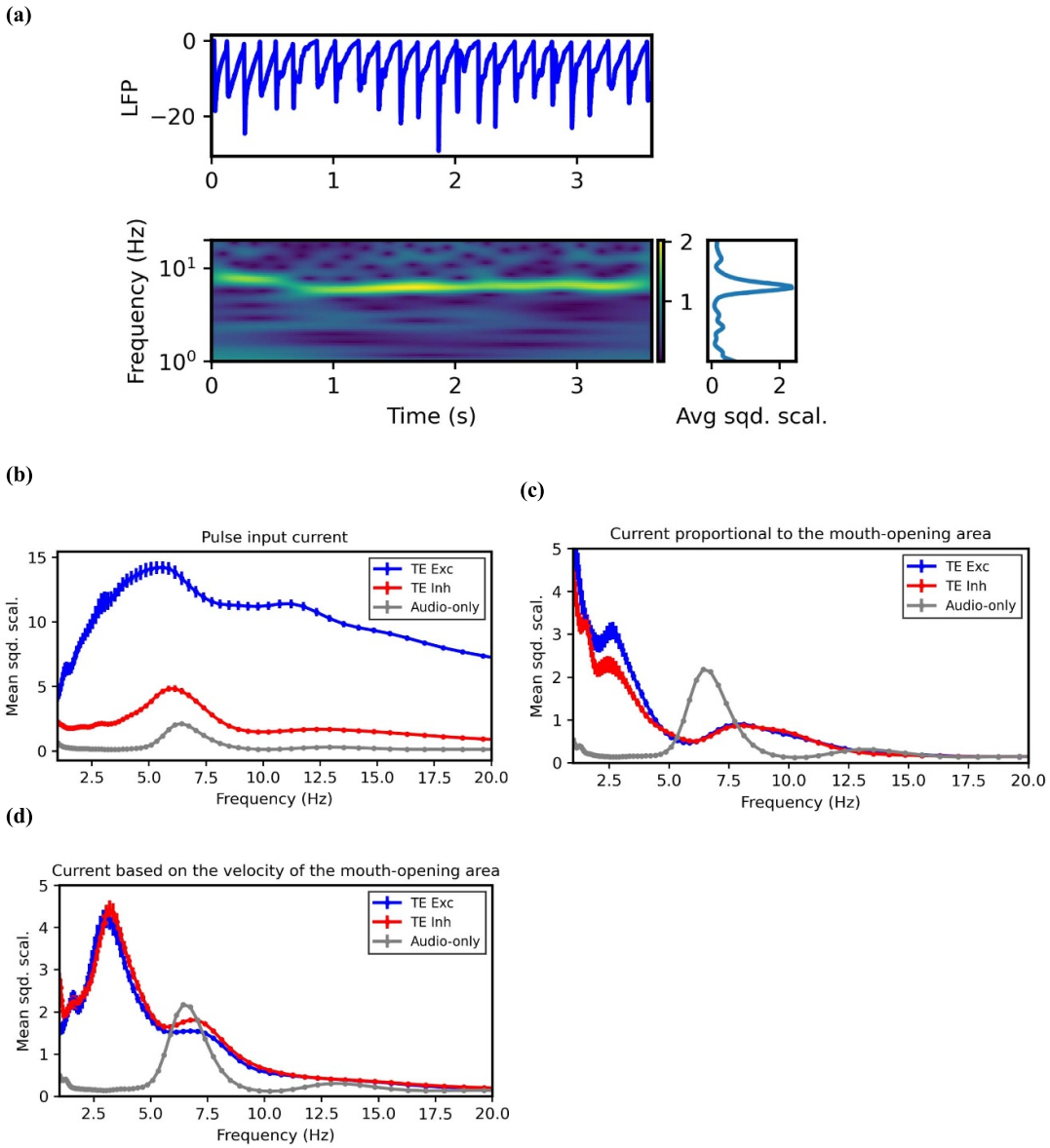

Figure 5. The mean squared scalogram for different visual currents to the excitatory neurons. (a) The spectrogram derived from an exemplary LFPs and the corresponding squared scalogram as a function of frequency. (b) The squared scalograms for three different conditions: no visual input (green), an excitatory pulse input (blue) and inhibitory pulse inputs (red). The excitatory inputs (blue) were added at 25 ms after the syllable onsets whereas the inhibitory input (red) was presented 125 ms before the syllable onset. (c) The squared scalograms for currents based on the mouth-opening area. The excitatory input (blue) preceded the auditory signal by 125 ms, whereas the inhibitory input (red) had a delay of 50 ms. (d) The mean squared scalograms for currents based on the velocity of the mouth-opening area. The excitatory input (blue) preceded the auditory signal by 25 ms whereas the inhibitory current (red) was presented with a delay of 125 ms.

speech processing and generate testable hypotheses for experimental studies. The values proposed in table 1 are a set of 'default' parameters, which may be furthermore modified if required. These values were previously used to systematically explore speech-in-noise processing in the model and were found to be a good fit showing similar trends to psychometric curves of human speech-in-noise comprehension [49]. Due to the relatively small size and low computational complexity, the model allows us to quickly screen a large space of hyperparameters (as we did here) to predict the effects of different conditions on the model behaviour.

We investigated how the accuracy of this syllable parsing changed when an additional current was added that mimicked different aspects of an accompanying visual signal. In particular, we added three different types of visual input currents to the network: a pulse current, a current corresponding to the mouth-opening area of the speaker, and a current corresponding to the velocity of the mouth-opening area.

We found that adding each of the three types of visual input currents could enhance as well as impede syllable parsing. However, syllable parsing was only affected when the current acted on the excitatory, but

not on the inhibitory neurons. We suppose that this is due to the recurrent inhibitory connections in the inhibitory population of the network which stunt the activity of the neurons in the presence of an external input current.

In the case of the pulse current, we observed that the parsing score as a function of the time lag exhibited some periodicity with a time period of 150 ms, which corresponded roughly to the time period of the theta oscillation. Furthermore, the dependency of the parsing score on the audiovisual time delay for the inhibitory pulse current was shifted with respect to the dependency for the excitatory current by about 100 ms, which suggested that the inhibitory currents inhibited the population which recovers after a theta cycle to be ready for the syllabic onset in the next theta cycle. Furthermore, in all these cases, we found that adding the visual current significantly improved the parsing score only at certain time lags of the visual current with respect to the auditory input current. This was because adding a visual input reset the phase of the theta LFP signal. The parsing score was greater if the LFP had an optimal phase at a syllable onset than when it had a nonoptimal phase.

Regarding the current based on the mouth-opening area, a significant improvement in the parsing score resulted when the visual current was inhibitory and delayed with respect to the auditory input by 50 ms, as well as when it was excitatory and had an advance of 125 ms. The current based on the velocity of the mouth area current could improve the syllable parsing when it was inhibitory and delayed by 150 ms.

When studying time lags between the auditory and the visual signal, we need to consider two components: the physical delay and the neural delay. The physical delay is the time difference between the onset of the audio signal and the visual signal, and the neural delay is the difference between the times it takes for the audio and visual input to reach the auditory cortex. The visual stimuli typically precede the auditory stimuli by about 100–300 ms, such as for the mouth movements of a speaker compared to the actual voice onset [8]. On the other hand, intercranial event-related potential (IERP) and voltage dye recordings indicate that the visual current arrives about 100 ms in the auditory cortex after the onset of the visual signal [32]. These two results indicate that the visual input may stimulate the auditory cortex about 100 ms before the auditory input does. Our finding that inhibitory pulse input preceding the syllabic onset by about 125–150 ms can enhance syllable parsing may therefore be particularly relevant.

We have studied both excitatory as well as inhibitory currents. However, a study in guinea pig auditory cortex using a voltage-sensitive dye showed inhibitory responses about 110 ms after the onset of the visual stimuli [20]. Another study in humans using fMRI found mainly suppressed activations in auditory cortices in response to visual stimulation [28]. Together with the likely earlier activation of the auditory cortex from visual rather than from auditory input, this suggests the enhancement of syllable parsing through an inhibitory, preceding pulse current may serve as a good model for understanding how visual inputs can enhance speech comprehension.

For the currents based on the mouth-opening area and on the corresponding velocity, the physical delay between the onset of the visual signal and the auditory signal is already incorporated in the signals. Therefore, we only need to account for the neural delay in the simulations. Considering a delay of 100 ms between the auditory and the visual signals, as suggested by experiments as described above, we find an improvement in the syllable parsing from a current based on the velocity but not from a current based on the actual mouth-opening area. The important feature might therefore be the mouth-opening area velocity rather than the mouth-opening area itself. Indeed, primary visual cortex is known to behave as an edge detector and a motion detector, responsible for computing the motion of objects across scenes [62].

When investigating which temporal lags, for a particular type of current, led to enhanced syllable parsing, we found lag regions with a width of about 50–100 ms. Future studies could try to design other currents with wider temporal regions to enhance syllable parsing as seen in experimental studies [63]. Furthermore, the simulations could be made for different playback rates of audiovisual input and compared with experimental input [64].

Because syllable parsing in humans cannot be measured behaviourally, our computational results cannot be compared directly to behavioural data. However, syllable parsing is required for syllable decoding, and the latter can be tested in experiments on speech comprehension. Moreover, the model predictions on neural activity could be tested in neuroimaging experiments. Experimental data could compare the power spectral density of EEG waves and the firing rates to see how they correspond to the simulations on the spectra of the LFP that we have done here to further shed light on the multisensory mechanism of audiovisual processing in the brain. Further studies could also tell us how visual speech affects the different oscillatory bands spatiotemporally across the auditory cortex and compare the results with the experimental data [65]. Integrating such neural data with behavioural data on speech comprehension in a computational model will further clarify the neural mechanisms of audiovisual speech processing.

Data availability statement

No new data were created or analysed in this study.

Acknowledgment

This research was supported by EPSRC grant EP/R032602/1 as well as by the U.S. Army through project 71931-LS-INT.

ORCID iDs

Anirudh Kulkarni https://orcid.org/0000-0002-1005-7671
Mikolaj Kegler https://orcid.org/0000-0003-3408-2588
Tobias Reichenbach https://orcid.org/0000-0003-3367-3511

References

[1] Alais D, Newell F N and Mamassian P 2010 Multisensory processing in review: From physiology to behaviour *Seeing & Perceiving* **23** 3–38
[2] Opoku-Baah C, Schoenhaut A M, Vassall S G, Tovar D A, Ramachandran R and Wallace M T 2021 Visual influences on auditory behavioral, neural, and perceptual processes: a review *JARO* **22** 365–86
[3] Benoit C, Mohamadi T and Kandel S 1994 Effects of phonetic context on audio-visual intelligibility of French *J. Speech Hear. Res.* **37** 1195–203
[4] Sumby W H and Pollack I 1954 Visual contribution to speech intelligibility in noise *J. Acoust. Soc. Am.* **26** 212–5
[5] Schroeder C E, Lakatos P, Kajikawa Y, Partan S and Puce A 2008 Neuronal oscillations and visual amplification of speech *Trends Cogn. Sci.* **12** 106–13
[6] Campbell R 2008 The processing of audio-visual speech: empirical and neural bases *Philos. Trans. R. Soc.* **363** 1001–10
[7] Munhall K G, Jones J A, Callan D E, Kuratate T and Vatikiotis-Bateson E 2004 Visual prosody and speech intelligibility: head movement improves auditory speech perception *Psychol. Sci.* **15** 133–7
[8] Chandrasekaran C, Trubanova A, Stillittano S, Caplier A and Ghazanfar A A 2009 The natural statistics of audiovisual speech *PLoS Comput. Biol.* **5** e1000436
[9] Felleman D J and van Essen D C 1991 Distributed hierarchical processing in the primate cerebral cortex *Cereb. Cortex* **1** 1–47
[10] Treisman A M and Gelade G 1980 A feature-integration theory of attention *Cogn. Psychol.* **12** 97–136
[11] Schroeder C E and Foxe J 2005 Multisensory contributions to low-level, 'unisensory' processing *Curr. Opin. Neurobiol.* **15** 454–8
[12] Kayser C, Petkov C I and Logothetis N K 2008 Visual modulation of neurons in auditory cortex *Cereb. Cortex* **18** 1560–74
[13] Kayser C, Petkov C I and Logothetis N K 2009 Multisensory interactions in primate auditory cortex: fMRI and electrophysiology *Hear. Res.* **258** 80–88
[14] King A J and Walker K M M 2012 Integrating information from different senses in the auditory cortex *Biol. Cybern.* **106** 617–25
[15] Ghazanfar A A, Chandrasekaran C and Logothetis N K 2008 Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys *J. Neurosci.* **28** 4457–69
[16] Bizley J K, Nodal F R, Bajo V M, Nelken I and King A J 2007 Physiological and anatomical evidence for multisensory interactions in auditory cortex *Cereb. Cortex* **17** 2172–89
[17] Atilgan H, Town S M, Wood K C, Jones G P, Maddox R K, Lee A K C and Bizley J K 2018 Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding *Neuron* **97** 640–55.e4
[18] Morrill R J and Hasenstaub A R 2018 Visual information present in infragranular layers of mouse auditory cortex *J. Neurosci.* **38** 2854–62
[19] Meng X, Kao J P Y, Lee H K and Kanold P O 2017 Intracortical circuits in thalamorecipient layers of auditory cortex refine after visual deprivation *eNeuro* **4** 1–11
[20] Kubota M, Sugimoto S, Hosokawa Y, Ojima H and Horikawa J 2017 Auditory-visual integration in fields of the auditory cortex *Hear. Res.* **346** 25–33
[21] Falchier A, Schroeder C E, Hackett T A, Lakatos P, Nascimento-Silva S, Ulbert I, Karmos G and Smiley J F 2010 Projection from visual areas V2 and prostriata to caudal auditory cortex in the monkey *Cereb. Cortex* **20** 1529–38
[22] Smiley J F and Falchier A 2009 Multisensory connections of monkey auditory cerebral cortex *Hear. Res.* **258** 37–46
[23] Bizley J K and King A J 2009 Visual influences on ferret auditory cortex *Hear. Res.* **258** 55–63
[24] Budinger E and Scheich H 2009 Anatomical connections suitable for the direct processing of neuronal information of different modalities via the rodent primary auditory cortex *Hear. Res.* **258** 16–27
[25] Cappe C and Barone P 2005 Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey *Eur. J. Neurosci.* **22** 2886–902
[26] Stehberg J, Stehberg J, Dang P T and Frostig R D 2014 Unimodal primary sensory cortices are directly connected by long-range horizontal projections in the rat sensory cortex *Front. Neuroanat.* **8** 1–19
[27] Pekkola J, Ojanen V, Autti T, Jääskeläinen I P, Möttönen R, Tarkiainen A and Sams M 2005 Primary auditory cortex activation by visual speech: an fMRI study at 3 T *Neuroreport* **16** 125–8
[28] Gau R, Bazin P L, Trampel R, Turner R and Noppeney U 2020 Resolving multisensory and attentional influences across cortical depth in sensory cortices *Elife* **9** e46856
[29] Zion Golumbic E, Cogan G B, Schroeder C E and Poeppel D 2013 Visual input enhances selective speech envelope tracking in auditory cortex at a 'Cocktail Party' *J. Neurosci.* **33** 1417–26
[30] Park H, Kayser C, Thut G and Gross J 2016 Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility *Elife* **5** 1–17
[31] Mégevand P, Mercier M R, Groppe D M, Golumbic E Z, Mesgarani N, Beauchamp M S, Schroeder C E and Mehta A D 2018 Phase resetting in human auditory cortex to visual speech *bioRxiv* 1–21
[32] Ferraro S *et al* 2020 Stereotactic electroencephalography in humans reveals multisensory signal in early visual and auditory cortices *Cortex* **126** 253–64
[33] Arnal L H and Giraud A L 2012 Cortical oscillations and sensory predictions *Trends Cogn. Sci.* **16** 390–8
[34] Gross J, Hoogenboom N, Thut G, Schyns P, Panzeri S, Belin P and Garrod S 2013 Speech rhythms and multiplexed oscillatory sensory coding in the human brain *PLoS Biol.* **11** e1001752
[35] Giraud A L, Kleinschmidt A, Poeppel D, Lund T E, Frackowiak R S J and Laufs H 2007 Endogenous cortical rhythms determine cerebral specialization for speech perception and production *Neuron* **56** 1127–34
[36] Etard O and Reichenbach T 2019 Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise *J. Neurosci.* **39** 5750–9
[37] Keshavarzi M, Kegler M, Kadir S and Reichenbach T 2020 Transcranial alternating current stimulation in the theta band but not in the delta band modulates the comprehension of naturalistic speech in noise *Neuroimage* **210** 116557
[38] Ding N and Simon J Z 2014 Cortical entrainment to continuous speech: functional roles and interpretations *Front. Hum. Neurosci.* **8** 1–7

[39] van Wassenhove V and Grzeczkowski L 2015 Visual-induced expectations modulate auditory cortical responses *Front. Neurosci.* **9** 1–10
[40] Simon D M and Wallace M T 2017 Rhythmic modulation of entrained auditory oscillations by visual inputs *Brain Topogr.* **30** 565–78
[41] Keil J and Senkowski D 2018 Neural oscillations orchestrate multisensory processing *Neuroscientist* **24** 609–26
[42] Luo H, Liu Z and Poeppel D 2010 Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation *PLoS Biol.* **8** 25–26
[43] Stanley D A, Falchier A Y, Pittman-Polletta B R, Lakatos P, Whittington M A, Schroeder C E and Kopell N J 2019 Flexible reset and entrainment of delta oscillations in primate primary auditory cortex: modeling and experiment *bioRxiv* 1–42
[44] Pittman-Polletta B R, Wang Y, Stanley D A, Schroeder C E, Whittington M A and Kopell N J 2020 Differential contributions of synaptic 2 and intrinsic inhibitory currents to 3 speech segmentation via flexible 4 phase-locking in neural oscillators *PLoS Comput. Biol.* **17** e1008783
[45] Kulkarni A, Ranft J and Hakim V 2020 Synchronization, stochasticity, and phase waves in neuronal networks with spatially-structured connectivity *Front. Comput. Neurosci.* **14** 569644
[46] Hyafil A, Fontolan L, Kabdebon C, Gutkin B and Giraud A L 2015 Speech encoding by coupled cortical theta and gamma oscillations *Elife* **4** e06213
[47] Jadi M P and Sejnowski T J 2014 Regulating cortical oscillations in an inhibition-stabilized network *Proc. IEEE* **102** 830–42
[48] Mazzoni A, Panzeri S, Logothetis N K and Brunel N 2008 Encoding of naturalistic stimuli by local field potential spectra in networks of excitatory and inhibitory neurons *PLoS Comput. Biol.* **4** e1000239
[49] Kegler M and Reichenbach T 2021 Modelling the effects of transcranial alternating current stimulation on the neural encoding of speech in noise *Neuroimage* **224** 117427
[50] Virtanen P *et al* 2020 SciPy 1.0: fundamental algorithms for scientific computing in Python *Nat. Methods* **17** 261–72
[51] Goodman D F M and Brette R 2009 The Brian simulator *Front. Neurosci.* **3** 192–7
[52] Garofolo J S *et al* 1990 *Acoustic-Phonetic Continuous Speech Corpus* (https://catalog.ldc.upenn.edu/LDC93S1)
[53] 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** 2421–4
[54] Chi T, Ru P and Shamma S A 2005 Multiresolution spectrotemporal analysis of complex sounds *J. Acoust. Soc. Am.* **118** 887–906
[55] Pillow J W, Shlens J, Paninski L, Sher A, Litke A M, Chichilnisky E J and Simoncelli E P 2008 Spatio-temporal correlations and visual signalling in a complete neuronal population *Nature* **454** 995–9
[56] Shi J V, Xu Y and Baraniuk R G 2014 Sparse Bilinear Logistic Regression (arXiv:1404.4104)
[57] Victor J D 2005 Spike train metrics *Curr. Opin. Neurobiol.* **15** 585–92
[58] Wilcoxon F 1946 Individual comparisons of grouped data by ranking methods *J. Econ. Entomol.* **39** 269
[59] Benjamini Y and Hochberg Y 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing *J. R. Stat. Soc. SeR.* B **57** 289–300
[60] Krischer L, Megies T, Barsch R, Beyreuther M, Lecocq T, Caudron C and Wassermann J 2015 ObsPy: a bridge for seismology into the scientific Python ecosystem *Comput. Sci. Discov.* **8** 0–17
[61] German-Sallo Z and German-Sallo M 2017 Multiscale analysing methods in electrocardiogram signal processing and interpretation *Procedia Eng.* **181** 583–7
[62] Pack C C, Born R T and Livingstone M S 2003 Two-dimensional substructure of stereo and motion interactions in macaque visual cortex *Neuron* **37** 525–35
[63] Grant K W and Greenberg S 2001 Speech intelligibility derived from asynchronous processing of auditory-visual information *Proc. Conf. Audit. Speech Process.* 132–7
[64] Brungart D S *et al* 2008 The effects of temporal asynchrony on the intelligibility of accelerated speech *Avsp* 19–24
[65] Ganesan K *et al* 2020 Visual speech differentially modulates beta, theta, and high gamma bands in auditory cortex *bioRxiv*