Audiovisual Matching in Speech and Nonspeech Sounds: A Neurodynamical Model

Marco Loh[1], Gabriele Schmid[2], Gustavo Deco[1,3], and Wolfram Ziegler[2]

Abstract

■ Audiovisual speech perception provides an opportunity to investigate the mechanisms underlying multimodal processing. By using nonspeech stimuli, it is possible to investigate the degree to which audiovisual processing is specific to the speech domain. It has been shown in a match-to-sample design that matching across modalities is more difficult in the nonspeech domain as compared to the speech domain. We constructed a biophysically realistic neural network model simulating this experimental evidence. We propose that a stronger connection between modalities in speech underlies the behavioral difference between the speech and the nonspeech domain. This could be the result of more extensive experience with speech stimuli. Because the match-to-sample paradigm does not allow us to draw conclusions concerning the integration of auditory and visual information, we also simulated two further conditions based on the same paradigm, which tested the integration of auditory and visual information within a single stimulus. New experimental data for these two conditions support the simulation results and suggest that audiovisual integration of discordant stimuli is stronger in speech than in nonspeech stimuli. According to the simulations, the connection strength between auditory and visual information, on the one hand, determines how well auditory information can be assigned to visual information, and on the other hand, it influences the magnitude of multimodal integration. ■

INTRODUCTION

In face-to-face communication, seeing the speaker's articulations enhances the intelligibility of speech. The ability of listeners to extract linguistic information from the movements of the speaker's jaw, lips, and tongue is called *speechreading*.

Although speechreading is not necessary to understand speech, it often has been shown that visual information from the speaker's face is beneficial for speech perception. This is true for normal hearing conditions (Campbell, 1996; Green, 1996; Summerfield, 1979; Erber, 1969), but especially also under conditions of a degraded acoustic signal, for instance, in noisy environments (Sumby & Pollack, 1954). Even when there is no acoustic signal at all, parts of an utterance can be understood by speechreading only (Summerfield, 1987; Erber, 1969). Moreover, visual information can alter auditory perception, as is demonstrated by the McGurk illusion (McGurk & MacDonald, 1976). In the McGurk experiment, an acoustic speech signal (e.g., /ba/) is dubbed onto a discordant visual speech signal (e.g., /ga/). The presentation of the discordant auditory and visual information leads to a perceptual illusion in which the listener perceives neither the acoustic nor the visual signal, but integrates the two channels to create a new percept (/da/). The McGurk illusion evidences that speech perception is a multimodal process.

What is the mechanism underlying audiovisual speech perception? By comparing speech gestures with nonlinguistic sound producing mouth movements, such as clicks or whistles, Schmid and Ziegler (2006) investigated whether the audiovisual interaction observed in speech processing is specific to the linguistic domain or if it represents an instance of a more general cross-modal processing ability. Research on audiovisual perception is often based on conflicting stimuli, like in the McGurk paradigm, with the aim of investigating the degree to which auditory and visual information are integrated. This approach requires labeling of the perceived stimulus, for instance, by phoneme categories. The method cannot be easily applied to nonspeech vocal tract gestures because there are no categories comparable to phonemes in the nonspeech domain.

To circumvent this problem, Schmid and Ziegler (2006) applied a match-to-sample design in which participants were asked to decide if two sequentially presented stimuli (auditory or visual) were the same or different. The main focus was on a cross-modal condition, in which participants had to decide if a heard sound matched a seen articulation. Although participants had no problem with the cross-modal matching of *speech* sounds, they showed high error rates in the *nonspeech* domain, indicating that the auditory-to-visual matching of speech gestures is not

[1] Universitat Pompeu Fabra, Barcelona, Spain, [2] City Hospital Bogenhausen, Munich, Germany, [3] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

based on implicit knowledge of the physical mechanisms of the movement-to-sound correspondence.

How can, in neural terms, this difference between speech and nonspeech conditions be understood? What is the critical difference between them? To investigate these issues, we propose a computational model to investigate the multimodal match-to-sample design. In particular, we target the connection strength between the two modalities as a possible source of the observed differences in the cross-modal matching of sounds. As basis for the match/nonmatch decision, we use the firing rates of the spiking neurons.

Because the match-to-sample paradigm does not allow us to draw conclusions about audiovisual integration mechanisms, we additionally simulate two further experimental conditions to target the integration of auditory and visual information. New behavioral data back up the predictive results of this simulation.

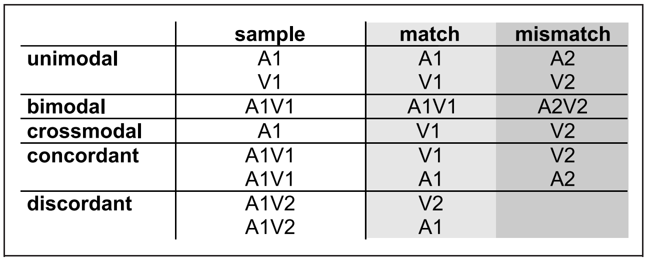
METHODS

Experimental Paradigms

Cross-modal Matching (Schmid & Ziegler, 2006)

The cross-modal matching paradigm was implemented in an experiment by Schmid and Ziegler (2006) and is the basis of the presented computational model. This experiment was based on pairs of speech (syllables) and nonspeech stimuli (sound-producing oral movements such as clicks, whistles, etc.). For the speech domain, we chose two pairs of reduplicated syllables. In one pair the distinctive feature was manner of articulation (/p/ vs. /f/) and in the other pair it was liprounding (/i/ vs. /y/). Because the visible gestures relating to these distinctions can be very short and because the critical mouth movement might interfere with a weakly specified onset position, the distinctive phonemes were embedded in syllables, and the syllables were produced in a reduplicated manner. Hence, for the manner distinction, we used the stimuli /apap/ versus /afaf/, and for the rounding distinction, we used the stimuli /nini/ and /nyny/. For the nonspeech domain, we chose a set of stimuli with comparable features. For the manner distinction, two bilabial stimuli were selected, namely, “kiss” and “whistle.” The rounding pair consisted of two alveolar clicks, one produced with spread lips and one with rounded lips. Analogous to the speech domain, the nonspeech stimuli were produced in a reduplicated manner as well.

The experimental task was to indicate, via keypress, whether two sequentially presented stimuli were equal (match-to-sample design). The two stimuli of a pair were presented with an ISI of 1 sec in order to prevent fusions or combinations to occur. Stimuli were presented in three modalities (see Figure 1): (1) unimodal visual (V), that is, both stimuli were mute video clips, and unimodal auditory (A), that is, both stimuli contained only the sound, with a gray square on the screen; (2) bimodal (AV), that is, both stimuli contained visual and auditory information; and (3) cross-modal (A × V), with the first stimulus being purely auditory and the second purely visual (Figure 1). Subjects were required to indicate, by pressing the appropriate key, if the two events they heard or saw were the same. In the cross-modal condition (A × V), participants had to decide whether the first stimulus (auditory) matched the second stimulus (visual). A total of 14 neurologically healthy subjects [9 men, 5 women; age (median) = 53 years, range = 34–67 years] took part in the study. For the analysis in this article, the unimodal visual and unimodal auditory conditions of this experiment were combined because we do not distinguish between modalities in the model presented here.



	sample	match	mismatch
unimodal	A1	A1	A2
	V1	V1	V2
bimodal	A1V1	A1V1	A2V2
crossmodal	A1	V1	V2
concordant	A1V1	V1	V2
	A1V1	A1	A2
discordant	A1V2	V2	
	A1V2	A1	

Figure 1. Experimental conditions of the discrimination task with 1-sec ISI between the sample and match period. The task contained five conditions: bimodal, unimodal, and cross-modal (see Schmid & Ziegler, 2006) and two new conditions: concordant and discordant (see text for details).

Audiovisual Integration

To confirm the predictions of the modeling work presented in this article, we studied two conditions related to audiovisual integration: concordant and discordant conditions (Figure 1). In these two conditions, the sample stimulus was presented bimodally and the second stimulus unimodally (either auditory or visual). The sample stimulus could either be concordant (matching auditory and visual information) or discordant (nonmatching auditory and visual information). In the discordant condition, there were no mismatch trials because the answer could be interpreted in two ways. If the subjects indicate a mismatch between sample and match stimulus, this can be because the mismatch was recognized in one domain or because the discordant sample was integrated (such as in the McGurk illusion), and thus, did not match the match stimulus. Because the match stimulus always corresponds to the sample stimulus in one domain, only the successful integration in discordant stimuli should lead to mismatch answers. Thus, we cannot determine in the discordant mismatch condition if integration took place or not (which was the aim of the study). This design allowed us to test integration effects not only in the speech but also in the nonspeech domain, as no explicit labeling of the items is required.

The stimuli were comparable to those from the cross-modal matching paradigm. Hence, the speech stimuli were two-syllabic nonsense items (e.g., /aba/, /aga/, /afa/). The nonspeech items were different types of labial and lingual sound-producing oral gestures, as for example, bilabial trill, bilabial click, alveolar click, and lateral click. A trained female speaker produced the stimuli at a moderate production rate. The stimuli were videotaped in a frontal view of the whole face including a part of the neck. To test audiovisual integration, we constructed discordant (McGurk-like) stimuli by suitable combinations of different speech stimuli, as for instance, auditory /aba/ and visual /aga/, or comparable combinations of nonspeech items, namely, auditory "bilabial click" with visual "alveolar click."

We tested 10 neurologically healthy subjects [5 women, 5 men; age (median) = 32 years, range = 23–53 years] on these two conditions. These new data have not been published yet.

Neurodynamical Modeling

Neural Network Model

We constructed a neural network model consisting of two modules, one for the auditory and one for the visual modality (Figure 2). They were envisioned to be selectively activated by the incoming stimuli, according to the experimental paradigm. As the neural basis of each module, we used a standard recurrent network model (Brunel & Wang, 2001), which has also been used to describe decision-making and other processes (Deco, Perez-Sanagustin, de Lafuente, & Romo, 2007; Deco & Rolls, 2006; Loh & Deco, 2005). We will first describe the neural level of the model and then present the network architecture.

The neural correlate has already been described in detail elsewhere (Loh & Deco, 2005; Deco & Rolls, 2003; Brunel & Wang, 2001). The neurons are represented by a leaky integrate-and-fire model:

$$C_{\mathrm{m}} \frac{\mathrm{d}V(t)}{\mathrm{d}t} = -g_{\mathrm{m}}(V(t) - V_{\mathrm{L}}) - I_{\mathrm{syn}}(t) \qquad (1)$$

where $V(t)$ is the membrane potential, C_{m} the membrane capacitance, g_{m} the leak conductance, and V_{L} the resting potential. Each module is made up of 400 excitatory and 100 inhibitory neurons. The synaptic input I_{syn} is made up of four parts. An external excitatory input via AMPA (alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid)-type synapses to the network model and recurrent input from the other neurons of the network. The latter consists of AMPA, NMDA (n-methyl-d-aspartate), and GABA (gamma-aminobutyric acid) currents. Thus, I_{syn} reads

$$I_{\mathrm{syn}}(t) = I_{\mathrm{AMPA,ext}}(t) + I_{\mathrm{AMPA,rec}}(t) + I_{\mathrm{NMDA,rec}}(t) + I_{\mathrm{GABA}}(t) \qquad (2)$$

The asynchronous external input $I_{\mathrm{AMPA,ext}}(t)$ can be viewed as originating from 800 external neurons firing at an average rate of $\nu_{\mathrm{ext}} = 3$ Hz per neuron, and thus, has a rate of 2.4 kHz in total. The neural network is fully connected and the recurrent input currents are summed over all neurons taking the connection weights into account. The synaptic dynamic is described by an exponential decay upon the arrival of a spike for AMPA and GABA currents and an alpha function including a rise time and an extra voltage dependence for the NMDA current. The detailed mathematical description and parameters are provided in the Supplementary Material. The parameters of the integrate-and-fire neurons and the synaptic channels for AMPA, NMDA, and GABA have been chosen according to biological data.

On the network level, the structure is composed of two modules. Each module contains two selective excitatory populations (A1, A2/V1, V2) (Figure 2), a nonselective excitatory population, and an inhibitory population. Each selective population contains 60 neurons, the nonselective population 280 neurons, and the inhibitory population 100 neurons. The connection weights between the neurons of the same selective population are referred to as intrapool connection strength w_+ (Figure 2). The other connection to the selective populations is w_- (also from the nonselective pool). The connection strength w_- is calculated so that the average connection strength input to a neuron equals 1 [$fw_+ + (1 - f)w_- = 1$, where $f = 0.15$ is the fraction of the number of neurons in a selective pool with respect to all excitatory neurons]. The excitatory con-

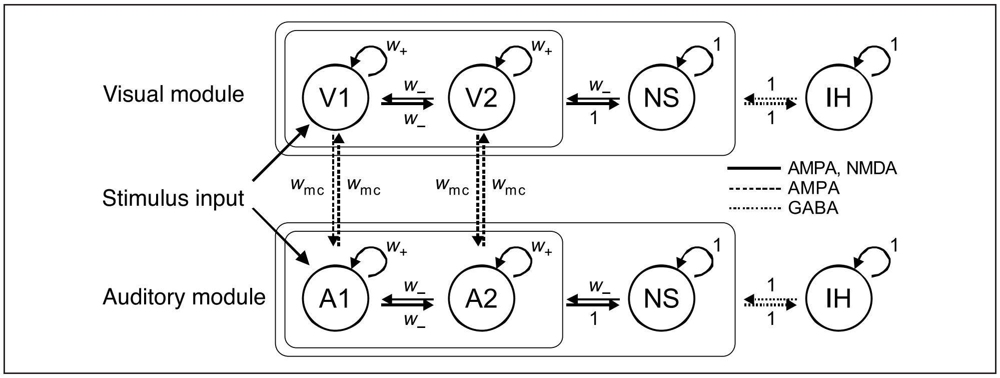


Figure 2. The cortical network model. The model contains two decision-making modules, one for each modality. Each module is made up of 400 excitatory and 100 inhibitory neurons. A detailed description of the model is found in the Supplementary Material.

nections are made up of AMPA and NMDA synaptic currents. All inhibitory GABA connections of the neurons in the inhibitory pool are set to the value 1 both to themselves and to all other neurons. The complete connection matrices are given in the Supplementary Material.

The two modules are envisioned to represent the auditory and visual decision-making area which are connected by a connection strength w_{mc} (Figure 2). For simplicity reasons, we implemented this connection only with AMPA synaptic currents. We hypothesize that speech and nonspeech domains might be characterized by different intermodule connection strengths.

Simulations

To analyze the network, we used spiking simulations. Spiking simulations calculate the dynamics of every neuron in time and yield a detailed temporal evolution of the system including fluctuations.

In our simulations, we ran the network for the sample, delay, and match period. The simulation protocol was as follows (Figure 3): The simulation protocol encompassed an overall period of 4 sec simulation time. The first stimulus (e.g., V1) was presented for 1 sec by increasing the external input to that pool selectively. Afterward, we simulated the ISI by running the network for another second without any external input. The matching stimulus was also presented for 1 sec followed by a 1-sec delay period.

The pools were simulated corresponding to the conditions depicted in Figure 1 and with equal numbers of match/mismatch pairs. For example, in the unimodal condition, we simulated two cases: match (V1–V1) and mismatch (V1–V2). Because the network is symmetric, we did not distinguish between the visual and the auditory regime. We simulated five conditions: unimodal, bimodal, cross-modal, concordant, and discordant, analogously to the behavioral experiments. For the discordant condition, we only simulated match trials because mismatch trials were omitted in the behavioral experiments (see above).

To implement a measure for matching stimuli in the paradigm, we compared the end of the delay period and the end of the match period (as indicated by the gray bars in Figure 3). We counted a trial as a match when the presentation of the match stimulus slightly increased the firing rate of the pool associated with the sample stimulus, that is, if the average firing rate of period P2 was higher than the one of period P1 (Figure 3). The selective pool should already be activated at a high level due to the presentation of the sample. We counted a trial as a mismatch if the firing rate in period P2 was lower than in period P1 or the difference between the two periods was very high (above 50 Hz). This indicates that the pool associated with the sample stimulus was not activated before and, therefore, no matching could occur. This might occur due to an altered perception (such as in the integration paradigm). If a mismatch stimulus is presented, the activity will decrease slightly because activation of another population will cause an indirect inhibition. We simulated for each condition 100 match and 100 mismatch trials.

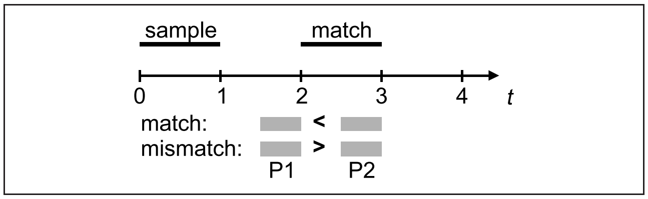


Figure 3. Simulation protocol. The simulation ran for 4 sec. The sample stimulus was presented between 0 and 1 sec, the match stimulus between 2 and 3 sec. To decide between match and nonmatch, we compared the average firing rate of 1.5–2 sec (P1) and 2.5–3 sec (P2). If the firing rate was higher in the second period, we counted it as a match, unless the difference was very high (above 50 Hz).

RESULTS

Behavioral Results

In the following we first will briefly describe the results of Schmid and Ziegler (2006), before we present the new behavioral results of the integration paradigm.

Figure 4A shows mean error rates for the different conditions of the behavioral study. In the unimodal conditions (results for the auditory and the visual conditions are collapsed) and the bimodal condition, error rates for nonspeech and speech stimuli were largely comparable. However, in the cross-modal processing, a significant difference between speech and nonspeech oral gestures was found: Although participants obviously hardly had any problems deciding if a heard speech sound matched a seen speech gesture (3.9% errors), they showed high error rates in the corresponding nonspeech task (23.5% errors). Hence, the most pre-eminent result of this experiment was a notable deterioration in the cross-modal, nonspeech condition.

Beyond these experiments, two new conditions were introduced to investigate the integration of auditory and visual information (cf. Figure 1). These were first investigated in simulations and then confirmed by the experimental data. In a *concordant* condition, in which a natural bimodal sample stimulus was compared to a unimodal match stimulus, low error rates occurred (speech: 2.9% errors, nonspeech: 1.7%; reaction times: speech: 1732 msec [*SD* ±126 msec], nonspeech: 1812 msec [*SD* ±184 msec]). In the corresponding discordant condition, the visual and the auditory components of the bimodal sample stimulus represented different sound/gesture categories. In this condition, mismatch responses (which are referred to as "errors" here) indicate successful integration of auditory and visual information within the sample stimulus. Although in the nonspeech domain 9.6% mismatch responses occurred, the rate of mismatch decisions in the speech domain amounted to 29.2% (see Figure 4A). These are significantly different (Wilcoxon signed-rank test, $p < .01$). The reaction

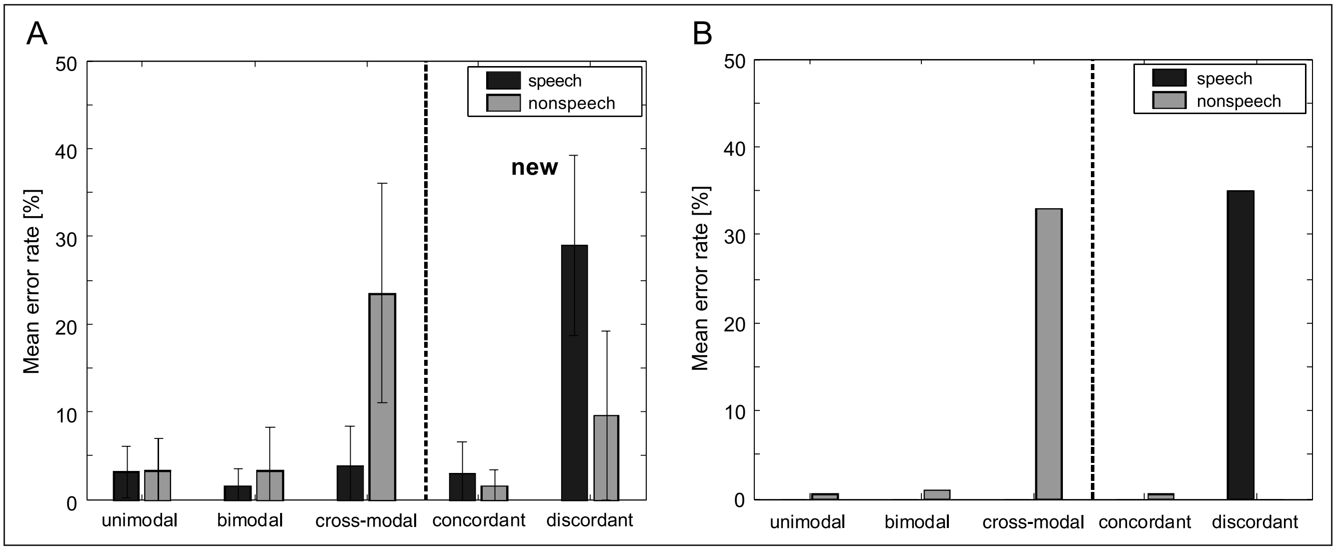


Figure 4. Mean error rates (%) of the match-to-sample paradigm. (A) Experimental data for both the speech and nonspeech domains for the unimodal and cross-modal case (Schmid & Ziegler, 2006) and new data on two integration paradigms (discordant/concordant). The error bars denote the standard deviation. The error rates of the discordant speech and nonspeech conditions are significantly different (Wilcoxon signed-rank test, $p < .01$). (B) Simulations using the parameters $w_+ = 1.6$, input = 150 Hz (on top of the external background rate of 2.4 kHz), and connection strength $w_{\mathrm{mc}} = 0.1/0.55$ for the nonspeech/speech domain, respectively.

times were 1905 msec [*SD* ±170 msec] for the speech condition and 1908 msec [*SD* ±194 msec] for the nonspeech condition.

Simulation Results

In the modeling work we hypothesize a mechanism based on a neurodynamical model, which could explain the experimental patterns described above. In particular, we suggest that the intermodule connection strength between the auditory and visual populations plays a key role.

There are three parameters in the model: the pool cohesion w_+ (which determines the connection strength within a pool), the input strength of the sample and match stimuli, and the intermodule connection strength w_{mc}. We fixed the pool cohesion and the input strength so that the model yielded bistable properties ($w_+ = 1.6$, input = 150 Hz). This is necessary to distinguish between the two possible stimuli. In the following, we discuss the influence of the remaining connection parameter w_{mc}, which could account for the performance differences in the speech versus nonspeech domain. Figure 5 shows the influence of intermodule connection strength on error rates in the different experimental conditions. Although the unimodal, bimodal, and concordant conditions are rather insensitive to changes in intermodule connection strength, error rates in both the cross-modal and the discordant conditions are modulated significantly: Whereas the cross-modal condition demonstrates high mean error rates at low connection strengths and improves with higher values, the discordant condition shows the opposite pattern, with increasing error rates at higher connection strength values. We suggest that the speech and nonspeech domain might work in dynamical regimes with high or low connection strengths, respectively. The underlying idea is that the greater experience with speech stimuli might have built up higher connection strength due to Hebbian learning mechanisms as compared to nonspeech stimuli.

Figure 4B illustrates the performance of the model for two different intermodule connection strengths ($w_{\mathrm{mc}} = 0.1$ vs. $w_{\mathrm{mc}} = 0.55$). Apparently, the model reproduces

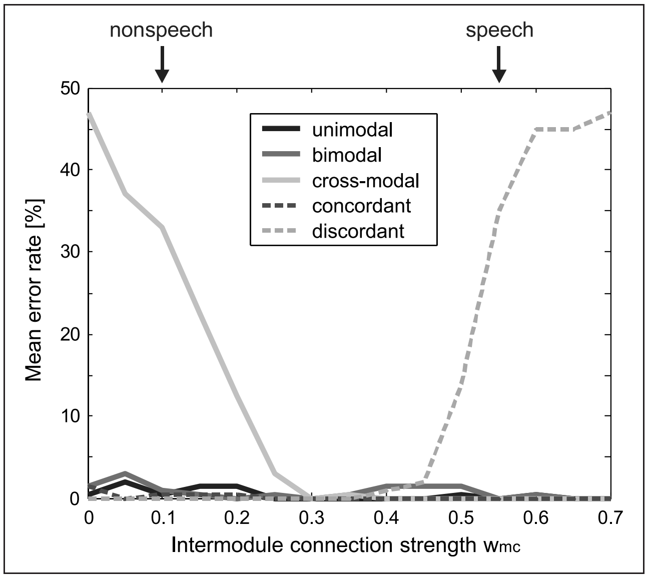


Figure 5. Mean error rates for simulations using the parameters $w_+ = 1.6$, input = 150 Hz (on top of the external background rate of 2.4 kHz), and varying connection strength w_{mc}. The different connection strengths might result due to varying experiences with stimuli. The speech/nonspeech conditions are at intermodule connection strength $w_{\mathrm{mc}} = 0.1/0.55$.

the experimental data shown in Figure 4A well if these two coefficients are considered to represent audiovisual connection strengths in the nonspeech (w_{mc} = 0.1) and the speech domains (w_{mc} = 0.55), respectively.

The unimodal and the bimodal conditions, along with the concordant condition, were characterized by low error rates in both the model and the behavioral data. In the cross-modal condition, the nonspeech task resulted in higher error rates than the speech task. Different intermodule connection strengths can account for the speech–nonspeech differences in cross-modal matching. Performance is better in the speech domain because the stronger connection between the two modules guarantees more efficient comparison of the two stimuli presented in different modalities.

In the discordant condition, the interference between the visual and auditory modality during sample presentation reduces response accuracy. This means that higher error rates occur due to successful integration of auditory and visual information, in this case of conflicting information which cannot be integrated to a coherent percept. The model generates more errors in the speech than in the nonspeech domain. This is because the high connection strength in the speech condition causes a stronger integration of the two conflicting sample stimuli, which reduces the probability of activating the populations associated with the sample stimulus correctly. If there is only little integration, like in the nonspeech condition, the unimodal match stimulus is predominantly compared with the corresponding channel of the sample stimulus, which results in higher match–response rates in the nonspeech as compared to the speech domain.

DISCUSSION

We constructed a model for a match-to-sample paradigm in the context of audiovisual perception of speech and nonspeech vocal tract gestures. Our major aim was to discuss the possibility that differences in cross-connection strengths between the auditory and visual modality might underlie the observed processing differences in the speech and nonspeech domain (Schmid & Ziegler, 2006). We showed in our model that a weaker intermodule connection for the nonspeech domain might be responsible for the differential processing in the two domains. Here we concentrated on the decreased performance in the cross-modal matching of speech as compared to nonspeech stimuli, and the higher rate of integration responses in the speech domain as compared to the nonspeech domain. A difference in intermodule connection strength can be seen as a candidate for these experimental effects. Due to perceptual learning and behavioral relevance, a stronger link between vision and sound may be established for speech than for novel nonspeech sounds. Changes in the cross-connection strength have proven to be sufficient to show the main effect in the experimental data, and thus, we kept the model as simple as possible. Top–down effects caused, for instance, by lexical access can be excluded in our study because we just used phonemes and nonspeech sounds as experimental stimuli.

One of the limitations here is that we used a symmetric model in which we treated the visual and auditory domain equally. Because in the underlying behavioral experiment the two unimodal tasks yielded similar results and we did not focus on modality differences, this approach is a good approximation. Nevertheless, a more detailed account of the audiovisual processing of vocal tract gestures might address the domain-specific differences between auditory and visual processing of linguistically meaningful and meaningless mouth movements. For example, the behavioral experiments revealed an advantage of the auditory over the visual modality in the discrimination of speech sounds, but not in the nonspeech modality. This suggests that processing differences exist between the speech and nonspeech modalities, which go beyond the aspects of intermodule connection investigated in this model.

Furthermore, we modeled two new conditions which had not been part of the original experiment (Schmid & Ziegler, 2006). We included two conditions in which a bimodal stimulus was presented during the sample phase and a unimodal stimulus during the match phase, with the bimodal sample stimulus either being concordant or discordant. A concordant stimulus is a natural bimodal stimulus in which the visual and auditory information match. In the discordant condition, the simultaneous visual and auditory stimuli do not match, like in the McGurk paradigm (McGurk & MacDonald, 1976). In contrast to the matching experiment, in which the main focus was on comparisons between the sample and the match stimulus, within-stimulus characteristics gained importance in the integration paradigm. We showed that in the nonspeech domain discordant information is not integrated in the same way as in the speech domain. The higher level of integration in the speech domain due to a stronger intermodule connection strength causes the system to integrate contradictory information and, thereby, changes the identity of the stimulus. Our model suggests that, in such an experiment, the number of correct matches should be lower in the speech domain than in the nonspeech domain. This was confirmed afterward by the behavioral data, which showed error rates of 9.6% versus 29.2% for the nonspeech versus speech domain, respectively. The high error rate in the nonspeech domain in the experiment could be due to a general difficulty in processing this type of stimuli, especially when they are presented in a nonnatural, discordant manner. Here, the model showed no errors. Nevertheless, the crucial point was to capture the main factor, that is, the difference in the discordant condition, which is significant in both experiment and simulation.

We have also seen that all conditions yielded almost entirely accurate decisions over a broad range of intermodule connection strengths (w_{mc} between 0.3 and 0.4;

Figure 5). This might be due to the fact that we investigated two different processes in one model: cross-modal matching and multimodal integration. It is an open question if, indeed, the same neural circuitry is involved in these two processes, as we implicitly assumed in our model. If the integration process probed in our second experiment engages a different processing network, this may potentially lead to a shift in the curves and reduce the range of connectivity values associated with accurate responses in all conditions.

Where are the auditory and visual processing networks for vocal tract gestures located in the brain, and how are they connected? We will briefly introduce two views discussed in the literature: First, these networks could be seated in primary sensory areas, in which either visual or auditory information is processed. The connection between these areas might either be direct or indirect via secondary regions. Hadjikhani and Roland (1998) defined "cross-modal-specific areas as areas activated only when information coming from two or more different sensory modalities is compared." These are areas which are responsible for the direct cross-talk between the two regions. Hadjikhani and Roland specifically exclude the possibility of a local storage of cross-modal information which is accessed by primary sensory processes. In investigations of the visuotactile system, they showed that the insula–claustrum might be a region carrying out the cross-talk between modalities. Other brain areas have also been suggested to participate in cross-modal functions such as anterior cingulate cortex, inferior parietal lobe, and left dorsolateral prefrontal cortex (Banati, Goerres, Tjoa, Aggleton, & Grasby, 2000). In this context, the auditory and visual modules of the model could be implemented in the primary auditory and visual area, respectively.

A second idea concerning the cross-modal processing of auditory and visual information considers a specific role of the superior temporal sulcus (STS). In the context of speech processing, this region is activated by both auditory (Binder et al., 2000) and visual stimuli (Calvert & Campbell, 2003). Moreover, this region shows properties of a multimodal integration site (Callan et al., 2003). This suggests that this region contains networks which separately represent auditory and visual information, but also networks specialized for the processing of multimodal information (Callan et al., 2003; Calvert, Campbell, & Brammer, 2000). Thus, the model neurons would be seated in the STS in possibly regionally overlapping neural networks. Note that we did not explicitly assume that the modules need to be spatially separated. A functional separation within the same area seems, especially in the STS, equally plausible.

However, the two views sketched here need not be mutually exclusive. For the speech domain, we can assume that the superior temporal lobe plays an important role in multimodal processing. This is shown by a large number of neurophysiological studies, which back the responsiveness of superior temporal cortex to visual, auditory, and audiovisual speech stimuli. Beauchamp, Argall, Bodurka, Duyn, and Martin (2004) even propose that this area serves as a general association learning device. The degree to which auditory and visual information are associated through perceptual learning might therefore be a crucial factor. As shown by the behavioral data, auditory and visual information seem to be only weakly associated in the nonspeech stimuli. Thus, the matching of information across modalities might, in this case, only be possible by a cross-talk between sensory-specific areas.

On the modeling level, we used a model based on integrate-and-fire neurons, which has also been used for decision-making (Deco & Rolls, 2006) and perceptual processes (Deco et al., 2007). The source of the probabilistic choices is fluctuations caused by noise in the system. There are two sources of noise: the Poissonian spike trains which excite the model externally, and the intrinsic finite size effects due to the low number of neurons in the modules. Thus, the level of modeling, here integrate-and-fire neurons, is crucial for the spiking noise which causes the perceptual decision-making. For the match/nonmatch choices, we used an artificial measure which compares two time windows in the simulation and decides using the difference in firing rate, if the two stimuli match. A neural implementation of this decision task would require a more detailed modeling. The current measure is a phenomenological implementation. In tactile experiments, detailed neural data exist for discrimination tasks (Brody, Hernandez, Zainos, & Romo, 2003) and first modeling work has been done on this match-to-sample paradigm (Machens, Romo, & Brody, 2005). More refined models could build on this work in order to include an explicit discrimination mechanism to compare the two stimuli.

Synchronous neural firing in the gamma range (30–100 Hz) of neural assemblies has also been proposed to underlie the binding of perceptual objects in cortical areas (Singer, 2001). This might also be a possible mechanism of audiovisual perception. Indeed, Kaiser, Hertrich, Ackermann, and Lutzenberger (2006) found that oscillatory activity is modulated by audiovisual stimulation, including illusionary effects. As our network activity is operating in the asynchronous regime, we did not analyze these effects. However, the same framework was already used to assess the effects of attention on rate versus oscillatory modulation (Buehlman & Deco, 2008). These effects could be analyzed in future work by adding the proportion of the AMPA and NMDA currents as additional parameter in the modeling process.

Future work might also address the causes of the differences between the speech and the nonspeech domain. We implemented the idea that different connection strengths between the two modules could be responsible for the experimental findings. Perceptual learning might have built up a stronger connection between the visual and auditory modalities for stimuli, which have been used most intensively during development. Clearly, language is of key importance and language-related representations of stimuli might be more strongly connected between modalities to increase the probability of correct communication. To address language acquisition, it might be interesting to investigate which learning mechanisms are suitable to build up cross-modal associations and how they evolve over time. The underlying neural mechanism would be of a Hebbian-type, which increases the connection strength between neurons that are activated frequently in a correlated manner.

Acknowledgments

The study was supported by the Volkswagen Foundation. We thank Ralph G. Andrzejak, Henning Boecker, Georg Goldenberg, Josef P. Rauschecker, Valentin Riedl, and Afra Wohlschläger for discussion.

Reprint requests should be sent to Gustavo Deco, Universitat Pompeu Fabra, Roc Boronat, 138, 08018 Barcelona, Spain, or via e-mail: gustavo.deco@upf.edu.

REFERENCES

Banati, R. B., Goerres, G. W., Tjoa, C., Aggleton, J. P., & Grasby, P. (2000). The functional anatomy of visual–tactile integration in man: A study using positron emission tomography. *Neuropsychologia, 38,* 115–124.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. *Nature Neuroscience, 7,* 1190–1192.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., et al. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex, 10,* 512–528.

Brody, C. D., Hernandez, A., Zainos, A., & Romo, R. (2003). Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. *Cerebral Cortex, 13,* 1196–1207.

Brunel, N., & Wang, X. (2001). Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *Journal of Computational Neuroscience, 11,* 63–85.

Buehlman, A., & Deco, G. (2008). The neuronal basis of attention: Rate versus synchronization modulation. *Journal of Neuroscience, 28,* 7679–7686.

Callan, D. E., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *NeuroReport, 14,* 2213–2218.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience, 15,* 57–70.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10,* 649–657.

Campbell, R. (1996). Seeing brains reading speech: A review and speculations. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines* (pp. 115–133). Berlin: Springer.

Deco, G., Perez-Sanagustin, M., de Lafuente, V., & Romo, R. (2007). Perceptual detection as a dynamical bistability phenomenon: A neurocomputational correlate of sensation. *Proceedings of the National Academy of Sciences, U.S.A., 104,* 20073–20077.

Deco, G., & Rolls, E. T. (2003). Attention and working memory: A dynamical model of neuronal activity in the prefrontal cortex. *European Journal of Neuroscience, 18,* 2374–2390.

Deco, G., & Rolls, E. T. (2006). Decision-making and Weber's law: A neurophysiological model. *European Journal of Neuroscience, 24,* 901–916.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research, 12,* 423–425.

Green, M. F. (1996). What are the functional consequences of neurocognitive deficits in schizophrenia? *American Journal of Psychiatry, 153,* 321–330.

Hadjikhani, N., & Roland, P. E. (1998). Cross-modal transfer of information between the tactile and the visual representations in the human brain: A positron emission tomographic study. *Journal of Neuroscience, 18,* 1072–1084.

Kaiser, J., Hertrich, I., Ackermann, H., & Lutzenberger, W. (2006). Gamma-band activity over early sensory areas predicts detection of changes in audiovisual speech stimuli. *Neuroimage, 30,* 1376–1382.

Loh, M., & Deco, G. (2005). Cognitive flexibility and decision making in a model of conditional visuomotor associations. *European Journal of Neuroscience, 22,* 2927–2936.

Machens, C. K., Romo, R., & Brody, C. D. (2005). Flexible control of mutual inhibition: A neural model of two-interval discrimination. *Science, 307,* 1121–1124.

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

Schmid, G., & Ziegler, W. (2006). Audio-visual matching of speech and non-speech oral gestures in patients with aphasia and apraxia of speech. *Neuropsychologia, 44,* 546–555.

Singer, W. (2001). Consciousness and the binding problem. *Annals of the New York Academy of Sciences, 929,* 123–146.

Sumby, W., & Pollack, I. (1954). Use of visual information for phonetic perception. *Journal of the Acoustical Society of America, 26,* 212–215.

Summerfield, A. Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lipreading* (pp. 3–52). Hillsdale, NJ: Erlbaum.

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