CORTEX 68 (2015) 61–75

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Special issue: Research report

Prediction across sensory modalities: A neurocomputational model of the McGurk effect

Itsaso Olasagasti[*], *Sophie Bouton and Anne-Lise Giraud*

Department of Basic Neurosciences, University of Geneva, Geneva, Switzerland

ARTICLE INFO

Article history:
Received 30 October 2014
Reviewed 12 December 2014
Revised 17 February 2015
Accepted 14 April 2015
Action editor Alessandro Tavano
Published online 30 April 2015

Keywords:
Predictive coding
Audiovisual integration
Computational modeling
McGurk effect

ABSTRACT

The McGurk effect is a textbook illustration of the automaticity with which the human brain integrates audio-visual speech. It shows that even incongruent audiovisual (AV) speech stimuli can be combined into percepts that correspond neither to the auditory nor to the visual input, but to a mix of both. Typically, when presented with, e.g., visual /aga/ and acoustic /aba/ we perceive an illusory /ada/. In the inverse situation, however, when acoustic /aga/ is paired with visual /aba/, we perceive a *combination* of both stimuli, i.e., /abga/ or /agba/. Here we assessed the role of dynamic cross-modal predictions in the outcome of AV speech integration using a computational model that processes continuous audiovisual speech sensory inputs in a predictive coding framework. The model involves three processing levels: sensory units, units that encode the dynamics of stimuli, and multimodal recognition/identity units. The model exhibits a dynamic prediction behavior because evidence about speech tokens can be asynchronous across sensory modality, allowing for updating the activity of the recognition units from one modality while sending top–down predictions to the other modality. We explored the model's response to congruent and incongruent AV stimuli and found that, in the two-dimensional feature space spanned by the speech second formant and lip aperture, fusion stimuli are located in the neighborhood of congruent /ada/, which therefore provides a valid match. Conversely, stimuli that lead to combination percepts do not have a unique valid neighbor. In that case, acoustic and visual cues are both highly salient and generate conflicting predictions in the other modality that cannot be fused, forcing the elaboration of a combinatorial solution. We propose that dynamic predictive mechanisms play a decisive role in the dichotomous perception of incongruent audiovisual inputs.

1. Introduction

In face-to-face communication speech is perceived through the visual and the auditory modalities. Compared with pure acoustic stimuli, the presence of a congruent visual stimulus enhances accuracy and shortens reaction times (Giard & Peronnet, 1999; Van Wassenhove, Grant, & Poeppel, 2005), and this effect is maximal when acoustic stimuli are weak, noisy or degraded. The performance enhancement induced by

* *Corresponding author*. Department of Basic Neurosciences, University of Geneva, Biotech Campus, 9 Chemin des Mines, C.P. 87, 1211 Genève 20, Switzerland.
E-mail address: miren.olasagasti@unige.ch (I. Olasagasti).
http://dx.doi.org/10.1016/j.cortex.2015.04.008
0010-9452/

visual cues in speech-in-noise occurs largely because vision and audition offer complementary information about the stimulus; vision conveys the place of articulation, while audition primarily conveys voicing and manner (Summerfield, 1987), providing concurrent cues that are ultimately merged in a single representation. Although at speech onset visual speech cues precede acoustic cues by approximately 100 msec (Chandrasekharan et al., 2009), in connected speech acoustic cues can precede visual cues by as much as 40 msec (Schwartz & Savariaux, 2014). The temporal correlations between visual and acoustic cues in normal speech hence define a 200 msec temporal window of integration (Massaro & Cohen, 1993; Munhall, Gribble, Sacco, & Ward, 1996; Stevenson & Wallace, 2013), ranging from approximately 30 msec of visual lag to about 170 msec of visual lead (Van Wassenhove, Grant, & Poeppel, 2007).

Audiovisual integration in speech perception is so powerful that it occurs even when the acoustic and visual streams are discrepant as exemplified by the McGurk effect (McGurk & MacDonald, 1976). In their seminal paper McGurk and MacDonald showed that visual /ga/ paired with auditory /ba/ leads to /da/ responses, termed *fusion*, whereas the responses to the opposite pairing of visual /ba/ with auditory /ga/ contained *combination* responses such as /bga/. Qualitatively, fusion has been described as the synthetic process by which the brain constructs a percept that coincides neither with the visual nor the acoustic modality. Combination, on the other hand, is usually described as a failure to fuse the two modalities, which results in the concatenation of the acoustic and visual tokens.

Here, we assume that incongruent audiovisual tokens leading to fusion and those leading to combination are qualitatively different, when taking into account the reciprocal predictions that visual and auditory modalities provide each other. We demonstrate that the incongruent simultaneous presentation of visual /aga/ and acoustic /aba/ closely matches a congruent /ada/ presentation in a two-dimensional space formed by lip aperture and the second formant (F2). In this case, the integrated predictions from both modalities do not conflict strongly and are close to /ada/. Conversely, no such close single-consonant audiovisual match exists for combination stimuli, which are characterized by salient visual and acoustic information. In that case, each modality provides strong and contradictory information about the other modality by way of cross-modal predictions. We hence hypothesize that the failure to find a single consonant match results in a combinatorial multi-consonant solution.

To illustrate these prediction effects across sensory modalities we used a hierarchical predictive coding framework (Friston, Trujillo-Barreto, & Daunizeau, 2008). Predictive coding is an optimal inference framework based on the idea that the brain internalizes forward models (how world events lead to sensory consequences), and that what travels from the sensory periphery to the brain are prediction errors (Rao & Ballard, 1999). The presence of predictive mechanisms in auditory and audio-visual speech processing has been shown experimentally (Bendixen, Scharinger, Strauß, & Obleser, 2014; Gagnepain, Henson, & Davis, 2012; Peelle & Davis, 2012; Sohoglu, Peelle, Carlyon, & Davis, 2012; Van Wassenhove, 2013) and explored at the theoretical level (Yildiz, von Kriegstein, & Kiebel, 2013). The model we present involves predictive mechanisms in audio-visual speech synthesis and, unlike previous works (Bejjanki, Clayards, Knill, & Aslin, 2011; Magnotti & Beauchamp, 2014; Magnotti, Ma, & Beauchamp, 2013; Massaro, 1998; Omata & Mogi, 2008; Yildiz et al., 2013), takes into account the dynamic processing of both acoustic and visual information.

2. Materials and methods

2.1. *Predictive coding model of AV speech perception*

Perception results from the processing of sensory inputs through a hierarchy of brain structures, where stimuli are represented with increasing levels of abstraction through a process that uses statistical knowledge about the environment.

To simulate this process, predictive coding uses a generative model, which represents the hierarchical structure and statistics of the world, and relates sensory inputs to their external causes. The brain's task is to infer the causes that create the sensory input, and this is simulated by inverting the generative model. The inversion involves top–down predictions from the generative model and bottom–up prediction errors. We used the model inversion based on Dynamic Expectation Maximization (DEM) (Friston et al., 2008). DEM inverts dynamic hierarchical models with a message-passing scheme that minimizes prediction errors. Activity at any given level predicts activity at the lower level using the generative model. Top–down communication relays predictions from a given level to the level below. Discrepancies between predicted and actual activity generate a bottom–up signal representing prediction error (PE). The level above can then use the PE signal to update its state so that its prediction becomes more accurate and prediction error minimal.

To apply DEM to AV speech perception we built a hierarchical generative model connecting a single multimodal recognition level to two sensory input modalities, auditory and visual. Between the recognition level containing abstract representations of congruent /aba/, /ada/ and /aga/, and the sensory level representing lip aperture (visual cue) and F2 (acoustic cue), we introduced an intermediate level of sequence units that determined the timing and ordering of lip and F2 associated with each speech token.

Fig. 1 shows the three levels of the model together with sample dynamics when confronted with a congruent /ada/ stimulus. Units at the top level, when active, generate both acoustic and visual estimates in the lower levels. Each recognition unit at the top level is associated with one of the three AV tokens through a distinct pattern of lip motion and second formant modulation in time (Fig. 2B). These internalized patterns are part of the generative model; they represent the lip and F2 sensory modulations as a sequence of 18 values. Since the speech token approximately corresponds to 400 msec of speech input, 18 points correspond to a temporal precision of approximately 25 msec. For each of the 18 time points there is a corresponding unit in the sequence level.

The generative model drives the sensory estimates by providing a target pair of lip and F2 values to the sensory level

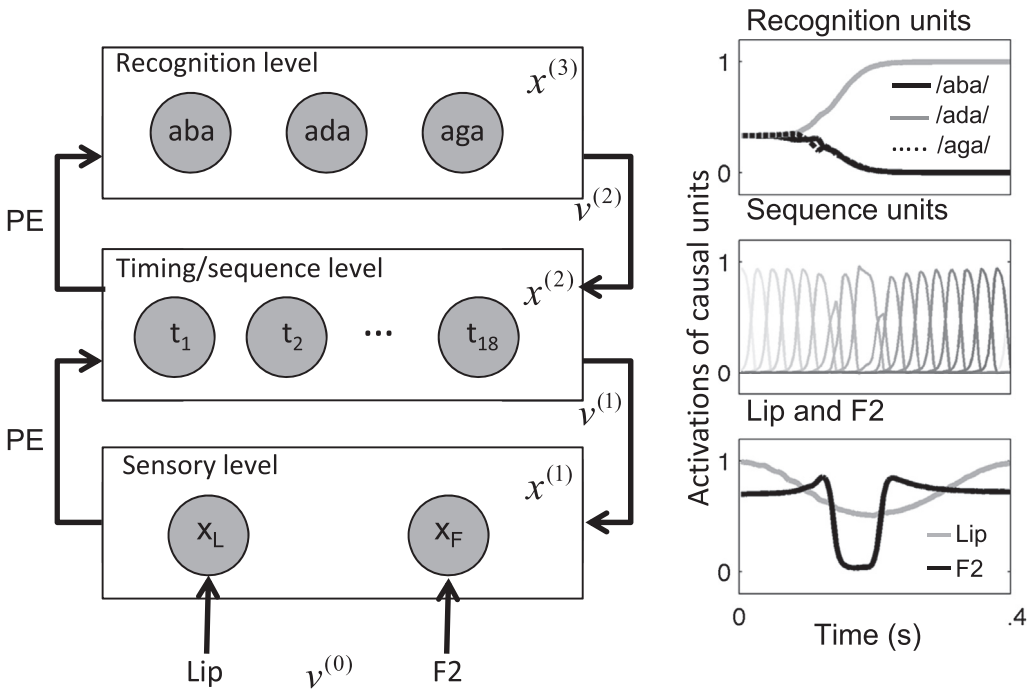

Fig. 1 – Left: Model schematics and flow of information. Top–down: current estimates of causal units; bottom–up: prediction errors. Prediction errors originate at the sensory level, in which lip and F2 predictions derived from the current state of the model is compared with incoming lip and F2 from the sensory periphery. Prediction errors are used to change the activity of the units in the model so that their generated lip and F2 predictions are closer to the sensory signals. Right: sample activity dynamics of causal units (v) at each level when the model recognizes congruent /ada/ input. Top: recognition units, middle: sequence units, bottom: sensory units. The plots show normalized activity for the top two levels and scaled versions of lip aperture and second formant for the audiovisual features.

at each time point ('Top–down input' in Fig. 3). Target lip and F2 values are determined by the internalized AV patterns and the instantaneous activation of recognition and sequence units. The dynamics at the sequence level are such that sequence units show transient sequential activation (Fig. 1, middle panel on the right). Consequently, for most of the time the input to the sensory level is dominated by the lip and F2 values corresponding to only one of the 18 points in the internalized patterns. The sequence level therefore selects the time point in the pattern and the recognition units select the appropriate pattern (/aba/, /ada/ or /aga/). At the start of the process all three recognition units are equally active and

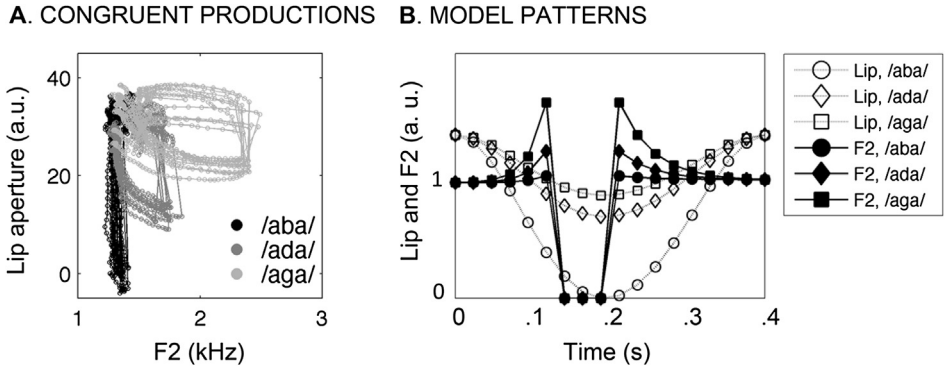

Fig. 2 – A. Lip aperture and second formant extracted from 10 productions of the three audiovisual speech tokens /aba/, /ada/and /aga/ from a single speaker. The data is represented as trajectories in the (F2, lip aperture) plane. The trajectory starts on the top branch and returns along the lower branch. The vertical lines signal the discontinuity in the second formant, which is not defined during the intervocalic interval. The figure shows that for a single speaker the trajectories are clearly separable. B. Based on the productions represented on the left panel we defined three pairs of lip aperture and second formant patterns used by the generative model to produce continuous top–down predictions for lip and formant values. The /aCa/ productions were represented in the model by lip aperture and F2 values sampled at 18 time points. The congruent productions generated by these patterns are shown in Fig. 3B.

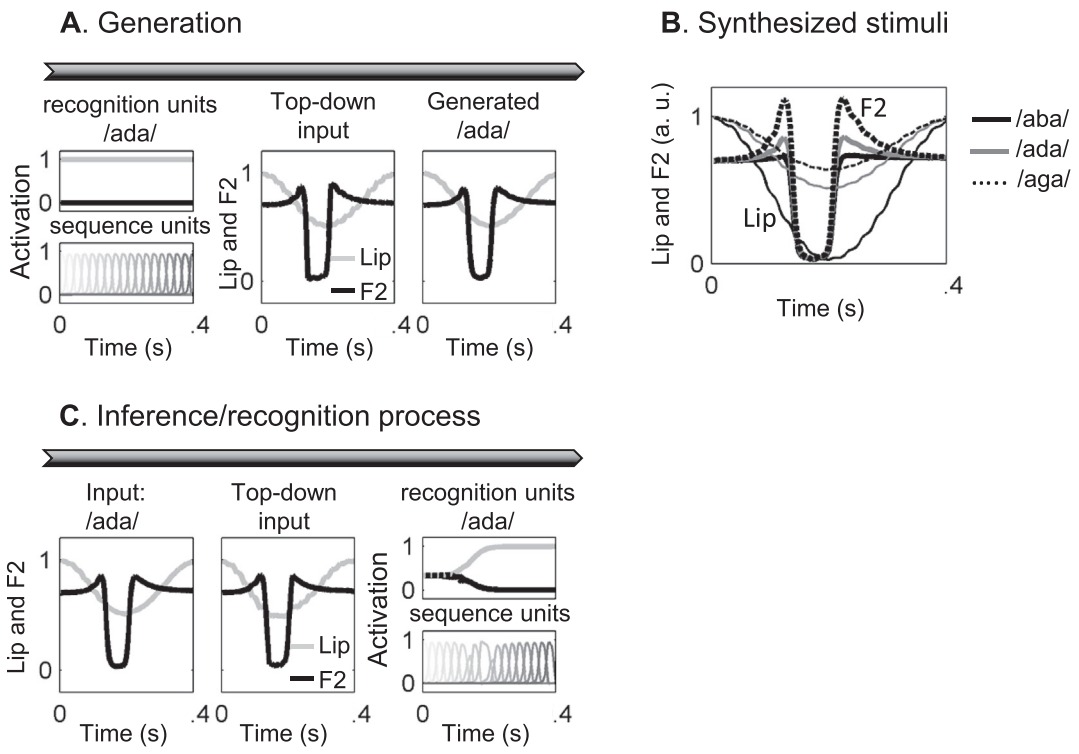

Fig. 3 – A. Generation of synthetic congruent audiovisual speech tokens. Left top and down panels represent the 2nd and 3rd levels of the model. The output of the /ada/ recognition unit is fixed to one (left panel, top) and the 18 sequence output units represented in different gray levels become sequentially activated (left panel, bottom). The two together, generate at each time step the driving input (middle) to the sensory level (right), which is a low-pass filtered version of the driving input. B. The three congruent patterns obtained by running the generative model using the patterns in Fig. 2B (/aba/ solid black lines, /ada/ solid grey lines, /aga/ black dashed lines; lip thin lines, F2 thick lines). C. Recognition of a congruent /ada/ token by the model. Left, the lip and F2 sensory inputs; middle, the driving input to the sensory level units consistent with the activity at the recognition level (right panel, top) and sequence units (right panel, bottom). In contrast to A, activity in the recognition units is driven by sensory inputs. Initially the three recognition units are similarly activated but as their associated patterns (B) separate from each other prediction error is minimized by increasingly activating the /ada/ recognition unit.

the target lip and F2 correspond to the mean of the three patterns. Since all patterns are indistinguishable at the beginning of the stimulus, there is no prediction error. As the modulation in the sensory input progresses, equal activation of the three recognition units generates a prediction error, which is minimized by increasing the activity of one of the recognition units. In the model recognition units have a time constant of ~ 200 msec, which allows them to integrate information, i.e., remember, over a period of time consistent with the known time window of integration in AV speech processing.

2.2. *Generative model equations*

The dynamics within each level are encoded by units/variables denoted by the letter *x*. Levels send information about their current state to the level below through causal/output units *v*. Discrepancies between the predictions coming from higher levels and the actual values are used in a message passing scheme to update state and causal units at each level so that prediction error is minimized. Both dynamic and causal variables are subject to random fluctuations, which, in the predictive coding implementation from Friston et al. (2008), can have temporal correlations.

The generative model has the general form:

$$\dot{x}^{(i)} = f\left(x^{(i)}, v^{(i)}\right) + \omega^{(i)}$$

$$v^{(i-1)} = g\left(x^{(i)}, v^{(i)}\right) + z^{(i)}$$

where (i) denotes the level. ω and z at each level represent random fluctuations; their covariance and precision (inverse covariance) play a critical role in the inference process. The inference process is driven by prediction errors (PE), the deviations from the generative model expectations: $\xi_x^{(i)} = \dot{x}^{(i)} - f\left(x^{(i)}, v^{(i)}\right)$ and $\xi_v^{(i)} = v^{(i-1)} - g\left(x^{(i)}, v^{(i)}\right)$. These travel from each level to the immediate level up the hierarchy (Fig. 1, left). How much each prediction error term contributes to the updating of *x* and *v* is related to their relative precisions; the smaller the fluctuation amplitudes of $\omega^{(i)}$ and $z^{(i)}$, the more the corresponding prediction error terms $\xi^{(i)}{}_x$ and $\xi^{(i)}{}_v$ will be weighted (Friston et al., 2008). As in standard Bayesian integration, the relative precisions will determine the relative

driving force that internal dynamics, top–down predictions and bottom–up prediction errors will have in the updating of states and estimates.

The left panel of Fig. 1 gives an overview of the model architecture with its three levels. Sensory inputs, one per modality, enter through the lower sensory level. $v^{(0)}$ represents the sensory inputs, which are compared with the top–down predictions $g(x^{(1)},v^{(1)})$. The resulting prediction error $v^{(0)}-g(x^{(1)},v^{(1)})$ is passed to the level above. The goal is then to change $v^{(1)}$ so that the prediction $g(x^{(1)},v^{(1)})$ is closer to the audiovisual input. Since $v^{(1)}$ is predicted by $x^{(2)}$ and $v^{(2)}$ through $g(x^{(2)},v^{(2)})$, any change in $v^{(1)}$ to improve the predictions for the sensory stimulus at the bottom level, will generate in turn changes on the higher level through the corresponding prediction error: $v^{(1)}-g(x^{(2)},v^{(2)})$. This is how the external sensory input can drive recognition at the top level (right panel of Fig. 1).

2.2.1. *Level 3 (top-level): consonant identity units*

There are 3 units in the top level (top left panel of Fig. 1). Each consonant from the set {b, d, g} is represented by one unit, which when active generates the corresponding /aCa/ F2 transitions and lip motion. There are state units x_i ($i = 1, \ldots,3$) with internal dynamics, and output units v_i ($i = 1, \ldots,3$). The output units transmit information to the lower levels; the output reaches level 1 through level 2.

$$\dot{x}_i^{(3)} = -a\left(x_i^{(3)} - c\right) - b\sum_{i\neq j} 1\Big/\left(1+\exp\left(-\alpha x_i^{(3)}\right)\right) + \omega_i^{(3)} \tag{1}$$

$$v_i^{(2)} = \exp\left(-x_i^{(3)}\right)\Big/\left(\sum_j \exp\left(-x_j^{(3)}\right)\right) + z_i^{(3)} \tag{2}$$

The dynamic equations for the states (x) contain a decay term (a: inverse time constant) and all-to-all inhibition (without self-inhibition). v is a normalized version of the activity of the state units. The top right panel of Fig. 1 shows the activity of the output units; initially the three units are similarly activated, but as sensory evidence arrives, the /ada/ unit becomes the one explaining best the sensory input. α and b are constants that amplify the winner-take-all component in the dynamics, while c sets the range of x. Parameter values were: $a = .015$, $b = .3$, $c = 2$, $\alpha = 3$.

During recognition dynamic units were initialized at the uniform fixed point obtained from Eq. (1). The equation $0 = -a(x-c)-2b/(1+\exp(-\alpha x))$, was solved numerically.

2.2.2. *Level 2 (middle level): sequence/timing units*

The dynamics of the states are such that only one unit is highly active, with the active unit changing in a predetermined order. A first set of state units (x) is connected so that they become active sequentially. A second set of state units (y) imposes normalization so that the output is bounded between zero and one (Yildiz et al., 2013). They represent timing units and they divide the entire duration into n_t intervals. We chose $n_t = 18$, so that each timing unit represented about 25 msec of speech.

$$\dot{x}_i^{(2)} = 2\left(-x_i^{(2)}\Big/18 - W_{ij}^{(2)}\left(1+\exp\left(-x_j^{(2)}\right)\right)^{-1} + 1\right) + \omega_{x,i}^{(2)} \tag{3}$$

$$\dot{y}_i^{(2)} = \left(\exp\left(x_i^{(2)}\right) - y_i^{(2)}\sum_j \exp\left(x_j^{(2)}\right)\right)\Big/ 2 + \omega_{y,i}^{(2)} \tag{4}$$

$$\boldsymbol{v}^{(1)} = \left(\max\left(\min\left(\boldsymbol{y}^{(2)},1\right),0\right),\boldsymbol{v}^{(2)}\right) + \boldsymbol{z}^{(2)} \equiv [\boldsymbol{y},\boldsymbol{u}]^T + \boldsymbol{z}^{(2)} \tag{5}$$

$$W_{i,i-1} = 0.5,\ W_{i,i} = 0,\ W_{i,i+1} = 1.5, \quad \text{otherwise } W_{i,j} = 1$$

$W_{i\ j}$ is a connectivity matrix from unit j to unit i and bold letters denote vectors. The output units send the value of the normalized states y together with the output received from level 3, $v^{(2)}$, which we rename as u. y carries information about an internally referenced time and u about the relative activation of recognition units. The middle panel of the right column in Fig. 1 shows the activity of the 18 y units in a simulation; each of them dominates the output at a specific time.

2.2.3. *Level 1 (low-level): sensory units*

The lowest level is made up of two units representing lip aperture and second formant (Bottom left panel in Fig. 1). We use a simple first order linear equation for the dynamics:

$$\dot{x}_i^{(1)} = 0.6\left(-x_i^{(1)} + \sum_{j,k} P_{ijk}u_j y_k\right) + \omega_i^{(1)},\ \ i = L,F\ \ j = 1,2,3\ \ k = 1,\cdots,18 \tag{6}$$

$$\boldsymbol{v}^{(0)} = [\boldsymbol{x}_L,\boldsymbol{x}_F]^T + \boldsymbol{z}^{(1)} \tag{7}$$

L and F stand for lip aperture (the visual cue) and second formant (the acoustic cue). P is a matrix that contains the time profiles for the whole /a-consonant-a/ (/aCa/) transition of the two sensory features (lip and second formant) associated with each of the 3 AV tokens (/aba/, /ada/, /aga/) sampled at $n_t = 18$ time points each. The way the patterns were defined is explained below under the heading Synthetic AV speech stimuli. This term changes, from moment to moment, the target value for the states x_i and consequently the internal estimate of the sensory inputs v. Matrix P represents a spatiotemporal pattern encoded by the identity/recognition units at the top level (Fig. 2B).

2.3. *Synthetic AV speech stimuli*

The input to be recognized by the model consisted of synthetic signals designed to capture the most salient features of audio and video signals of actual /aba/, /ada/ and /aga/ productions from a single speaker. Acoustic stimuli were represented by the second formant (F2), which was extracted from the sound waveform with Praat (Boersma & Weenink, 2014). We chose F2 because it can reliably be detected and distinguishes between /aba/ /ada/ and /aga/. Visual stimuli were represented by lip aperture, which was extracted from the video with custom scripts written in MATLAB (The Mathworks, Natick, Massachussetts, version 8.2.0.701).

Synthetic signals were obtained by running the generative model presented above. The critical quantities defining the characteristics of the acoustic and visual signals generated by

the model are the patterns encoded in matrix P of the first, sensory, level. As explained above, P encodes the lip and formant profiles of congruent AV speech tokens sampled at 18 consecutive time points.

To define reasonable values for P we selected about 400 msec of the audiovisual signal from the productions of a single speaker. This interval was bounded by the maximum lip aperture before and after the consonantal constriction. F2 was divided by 2000 Hz and lip aperture was normalized so that maximal aperture was 1. From the median time profile across the 10 productions of each token we extracted F2 values immediately before (F_{off}) an after (F_{on}) the intervocalic interval, approximate timing of intervocalic interval (t_{off} and t_{on}), as well as the amplitude of lip closure (L_c). The 18 values for the pattern were then chosen from the following parameterization, which captures the main F2 and lip aperture features extracted from the medians:

$$F(t) = \begin{cases} F_0 + (F_{off} - F_0)\exp(-(t_{off} - t)/\tau_{off}) & t < t_{off} \\ 0 & t_{off} < t < t_{on} \\ F_0 + (F_{on} - F_0)\exp(-(t - t_{on})/\tau_{on}) & t > t_{on} \end{cases} \quad (8)$$

$$L(t) = (1 - L_c[-\cos(2\pi t/T) - \cos(4\pi t/T)/20 + (19/20)]/2) \quad (9)$$

T is the total duration of the speech token. The values from the medians of the 10 productions of each of the three AV speech tokens were as follows: F_{on}: (.75, .9, 1.2) and L_c = (1, .5, .37) for /aba/, /ada/, /aga/ respectively. For simplicity we set $F_{off} = F_{on}$.

Unlike lip aperture, the second formant is not defined for the whole duration of the speech token since it is not present during the oral cavity constriction. This is reflected in an overall decrease in power during the intervocalic interval. We accounted for this fact indirectly by setting pattern values for $F2 = 0$ during this interval. Since all recognition (Level 3) units will generate the same formant value during this interval, there is no information about the speech token to be extracted. However, there is a strong cue about timing, which can bias/reset the activity of units at the sequencing/timing level. Thus our F2 can be regarded as a combined second formant and onset/offset signal. The AV signals produced by the generative part of the model presented above with these three patterns defined our synthetic congruent inputs. Thus we fixed the activity of the relevant recognition unit at the top level and saved the resulting lip and F2 outputs ($L(t)$, $F(t)$). The 400 msec of stimulus were represented by 120 time points, which corresponded to ~ 3msec time steps.

To simulate variability across and within speakers in AV speech production, we created stimuli with varying lip and F2 profiles in a 2D continuum by considering independent linear mixtures of /aba/ and /aga/ lip and F2 components:

$$L(t) = a_L L(t)_{aba} + (1 - a_L)L(t)_{aga} \quad (10)$$

$$F(t) = a_F F(t)_{aba} + (1 - a_F)F(t)_{aga} \quad (11)$$

$L(t)_{aba/aga}$ stands for the lip aperture profile, $F(t)_{aba/aga}$ for the second formant profile (from the congruent generation). In this parameterization the congruent tokens are represented by (a_L, a_F): /aba/= (1, 1), /aga/= (0, 0), /ada/= (.21, .67). The values corresponding to /ada/ were obtained by solving the equations derived by using the values for L_c and F_{on} given above (by definition $L_c = L(t_c)$ and $F_{on} = F(t_{on})$) with t_c the time of maximum lip closure and t_{on} the time of F2 onset after the intervocalic interval. $L_{ada}(t_c) = a_L L_{aba}(t_c)(1 - a_L) + L_{aga}(t_c) \Rightarrow a_L = (L_{ada}(t_c) - L_{aga}(t_c))/(L_{aba}(t_c) - L_{aga}(t_c))$

Similarly, $a_F = (F_{ada}(t_{on}) - F_{aga}(t_{on}))/(F_{aba}(t_{on}) - F_{aga}(t_{on}))$

2.4. Simulations

The simulations presented in this paper were performed using the dynamic expectation maximization tools of SPM8 (Wellcome Trust Center for Neuroimaging, UCL), running in MATLAB. The integration time step was the same as the time step of the stimulus (approximately 3 msec). Qualitatively similar results were obtained when running a control simulation with half the time step.

The fixed log precision parameters (log P), unless otherwise indicated, were as follows. Level 3, states and causes log $P = 12$. Level 2, states and causes log $P = 10$. Level 1, states and causes for lip aperture log $P = 5$, states and causes for F2 log $P = 8$.

The outcome of the inference process depends on precision values. The values used here were chosen to illustrate how cross-modal predictions can lead to fusion and combination stimuli. To reduce the number of free parameters we considered equal precision for causal and hidden states at each level. For noisy sensory inputs one should rather opt for a higher precision for hidden states than for causal states (e.g., Yildiz et al., 2013), but since we did not use noisy stimuli, this was not necessary.

Our goal was to find parameters that *i*) lead to recognition of congruent tokens and *ii*) result in similar effective strengths of visual and acoustic cues. The first requirement was met by setting higher precision values for higher levels in the hierarchy. To meet the second requirement we chose higher precision values for acoustic than visual cues. With equal precision for both sensory cues, the visual cue has a stronger impact on the outcome because it gives information at every time point, whereas acoustic information is not available in the intervocalic period. In Fig. 5 we illustrated how the inference outcome changes with the relative precision of acoustic and visual cues. Choosing more extreme precision ratios in favor of one of the sensory modalities leads to the expected result that both fusion and combination are biased towards the stronger modality (data not shown). Assuming that people may differ in the relative precision they assign to acoustic and visual cues, interindividual variability might explain the different incidence of combination and fusion responses in psychophysics experiments. Future work will explore more systematically the effect of precision and use behavioral data to constrain them as much as possible.

3. Results

3.1. Separability of audiovisual speech tokens

To represent the audiovisual tokens we chose a single feature per modality; the second formant of the spectrogram for the auditory stream and lip aperture for the visual stream. Fig. 2A

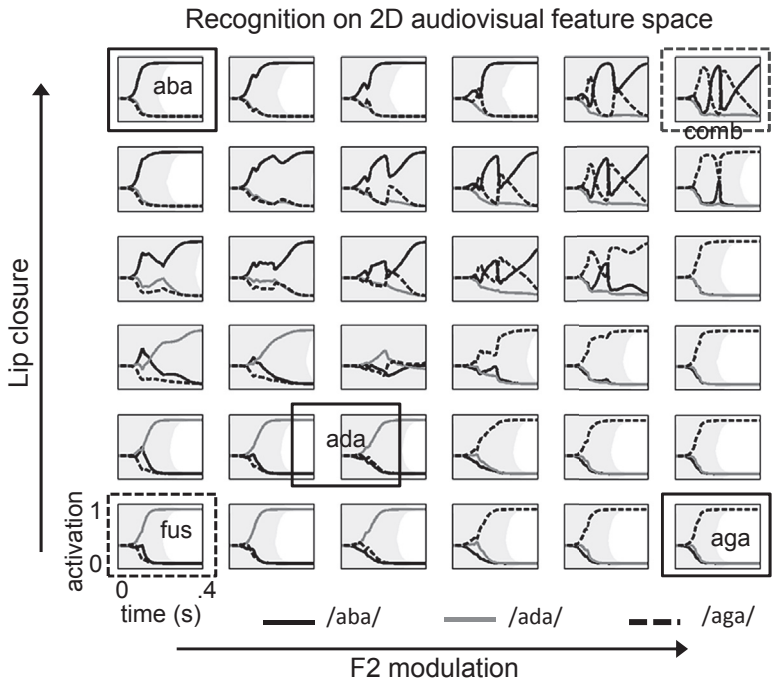

Fig. 4 – Recognition units in response to 36 combinations of lip aperture and F2. Each individual panel shows the activity of the recognition units as a function of time. Lip closure is constant along rows, with the top row having the maximal lip closure (corresponding to /aba/) and the bottom row minimal lip closure (/aga/). Columns are arranged according to the size of the formant transition, maximal for /aga/ (right column) and minimal for /aba/ (left column). Stereotypical congruent audiovisual tokens are marked with a solid black rectangle, the stereotypical fusion and combination stimuli with a dashed grey, respectively, black rectangle. The shaded regions reflect the 95% confidence intervals. Fusion stimuli are relatively close to a congruent /ada/; combination stimuli however (dashed grey square) are not close to a single consonant transition and no single recognition unit can appropriately account for the stimulus. Some features of the pattern of activations are rather robust to changes in relative precisions. The acoustic cue for /aga/ and the visual cue for /aba/ are particularly salient; this is reflected in the rapid rise of the activity related to /aga/ (dashed black) in all the stimuli of the right column, and the rapid rise of /aba/ activity (solid black) on the top row. When the activity of one unit predominates over the others, it is effectively making a prediction in time about the identity of the stimulus; the top–down component in the model sends an /aba/ respectively /aga/ prediction. When the predicted sensory input conflicts with the actual input (top right quadrant of the picture), prediction error leads to a reevaluation, the activity does not stabilize and both /aba/ and /aga/ units are strongly active at some point during the stimulus. In contrast, fusion stimuli (visual /aga/, auditory /aba/) are processed similarly to congruent /ada/.

shows 10 trajectories in the lip aperture and formant space, for each of the three tokens /aba/, /ada/, /aga/from a single speaker. Although additional features would be needed to further distinguish, for example, between the /bdg/ and /ptk/ consonants, this representation is sufficient for the current purpose. The three /aCa/ tokens have distinct trajectories and we created synthetic signals with the same qualitative features; duration of intervocalic interval, peak second formant value just before and after the occlusion, related drop in sound intensity, and size of lip aperture modulation (amount of lip closure).

3.2. *Generation and inference in a hierarchical predictive coding model of AV speech perception/categorization*

The hierarchical generative model was designed to explore the predictive action of the visual and acoustic modalities in AV speech processing. The relationship between the activity at the top/recognition level of the hierarchy and the bottom/sensory level qualitatively mimics how an actual speech token leads to the associated audiovisual percept. We will show that fusion stimuli are processed very similarly to congruent /ada/ because the prediction of the early activity is not contradicted by subsequent features in the stimuli. In contrast, the processing of combination stimuli leads to predictions early in the stimulus presentation that are subsequently contradicted by later features in the stimuli.

When the model is used for perceptual inference, if there is a discrepancy between sensory inputs and the estimate of the sensory inputs derived from the current state of dynamic and causal states in the model, error signals (PE) are passed to higher levels and are used to update the dynamic and output units in all levels so that prediction errors are minimized (Friston et al., 2008). Our basic approach to study the

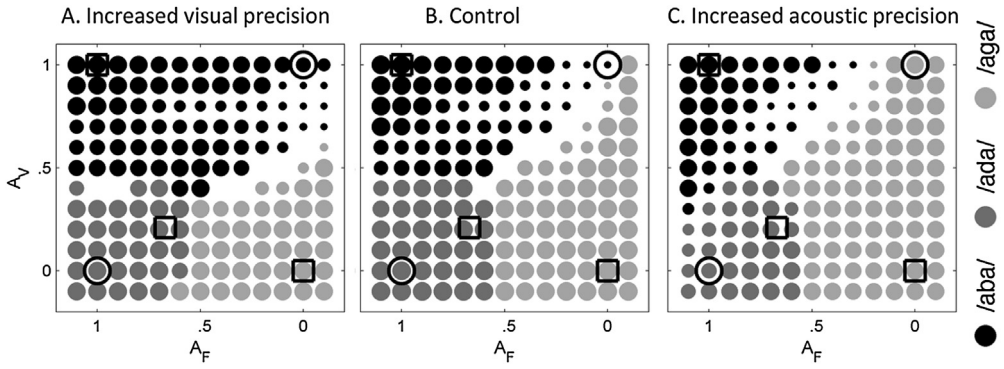

Fig. 5 – Output of the recognition process when the precision associated with either the visual cue (left panel) or the acoustic cue (right panel) is increased relative to our standard simulations (center panel, Fig. 4 and Fig. A.1). Input stimuli consisted of 12 × 13 pairs of lip aperture and F2 modulations obtained by varying the amount of lip closure and F2 modulation independently (see Methods for a detailed description). The position on the grid determines the relative distance from the /aba/ stimulus in the lip closure dimension (A_V) and F2 modulation dimension (A_F). The positions of the standard /aba/ ($A_V = 1$, $A_F = 1$), /ada/ ($A_V = .21$, $A_F = .67$) and /aga/ ($A_V = 0$, $A_F = 0$) are marked with a square, and those of the standard fusion ($A_V = 0$, $A_F = 1$) and combination stimuli ($A_V = 1$, $A_F = 0$) by a circle. At each point of the grid, the color of the filled circles represents the identity of the recognition unit with highest activity at the end of the process. The size of the filled circle is inversely related to the fluctuations in recognition unit activity (smaller circles related to larger cross-modal conflict). Only precisions for causal states at the bottom level differed across panels. A: log precision for lip causal state log $P(v_L) = 6$, and for F2 causal state log $P(v_F) = 7$. B: log precision for lip causal state log $P(v_L) = 5$, and for F2 causal state log $P(v_F) = 8$. C: log precision for lip causal state log $P(v_L) = 4$, and for F2 causal state log $P(v_F) = 9$.

qualitative features of prediction in multimodal processing of AV speech signals is to use a model that can robustly recognize congruent AV speech tokens and test its performance on non-congruent speech tokens to unveil the differences between fusion and combination responses.

Instead of building a model that learns appropriate patterns from congruent productions from an actual speaker, we considered a model that generates congruent AV speech signals, sharing the main features of recorded productions from a single speaker (see Methods). Model-synthesized speech signals (Fig. 3B) and their incongruent combinations were subsequently used as inputs to the model for identification/perceptual inference.

The generative part of the model is illustrated in the top row of Fig. 3, in which the activated unit encodes /ada/ ($v^{(2)}_{ada} = 1$). The top–down driving input to the first level dynamic units ($x^{(1)}_{L,F}$) is a weighted sum of the pattern values over time and over identity (Eq. (6), Fig. 1B). The instantaneous weights are given by the instantaneous activities of sequence output units ($v^{(1)}$) and recognition output units ($v^{(2)}$). During generation, only one of the three recognition output units is active and from the sequence output units one dominates over the others most of the time; therefore the input to the first level successively goes through the values encoded in the pattern at the 18 sampled time points (corresponding to a temporal precision of ~25msec) one after the other at the natural rhythm of the sequence units (which we fixed so that the whole 18 unit sequence would span 400 msec of AV speech).

Fig. 3B shows the three lip and F2 profiles obtained for the three congruent /aCa/ tokens by running the generative model with the patterns of Fig. 2B. We considered these profiles as the standard congruent sensory stimuli. During inference, the model is driven by sensory stimuli. In the simulation illustrated in Fig. 3C the inference model is confronted with one of the standard congruent sensory stimuli obtained with the generative model, in this case /ada/. Since the three congruent patterns (Fig. 2B) initially have the same values, at the very beginning none of the top-level recognition units is favored. However, as the separation of lip aperture and second formant amongst the three profiles increases (Fig. 3B) the /ada/ encoding unit gets increasingly activated and eventually dominates the output of the top level.

3.3. *Predictive contributions to the processing of incongruent AV speech tokens*

Finally the model defined by the three congruent /aba/, /ada/ and /aga/ recognition units was presented with parametric combinations of lip and F2 modulations. Synthetic stimuli were constructed by varying independently the degree of lip closure and peak F2 between their minimum and maximum values. Lip closure is minimal for /aga/ and maximal for /aba/, while the opposite is true of peak F2; maximal for /aga/ and minimal for /aba/ (Fig. 3B). This family of stimuli contained congruent-like stimuli, fusion-like stimuli and combination-like stimuli and allowed us to illustrate the qualitative difference between fusion and combination stimuli. The synthetic congruent stimuli captured the main features of lip and F2 profiles of a set of AV productions from a single speaker. By considering a two parametric family of lip and F2 pairings we could qualitatively mimic the variability across speakers and trial-to-trial productions.

Fig. 4 shows the activity of the three recognition units at Level 3 for 36 lip closure and F2 modulation pairs (Fig. A.1 shows the same simulation with a finer sampling of the lip closure and F2 space). The pairs closest to the three congruent AV tokens are marked with a black square. The response to a fusion-like stimulus (lip profile from /aga/ presented together with the F2 profile from /aba/) is surrounded by a solid grey square and that to a combination-like stimulus (lip profile from /aba/presented with the F2 from/aga/) by a dashed grey square. Importantly, the position on the grid of the congruent AV stimuli (black solid squares) illustrates that /aba/ and /ada/ are closer in the acoustic modality and /aga/ and /ada/ are closer in the visual modality. As a result, fusion stimuli are closer to a congruent stimulus than combination stimuli and, as is the case in the simulation of Fig. 4, the incongruence might not be detected and the fusion stimuli can be perceived as /ada/.

As is typical in optimal inference, the relative precision (inverse covariance of fluctuations) of the variables in the model plays a determinant role in the outcome of the inference process; quantities with higher precision (lower amplitude fluctuations) drive the inference process more strongly. In particular the relative precision of the two sensory modalities determines which modality is going to be weighed more, while relative precisions across levels in the hierarchy determine the relative weight of bottom–up sensory driven changes and of top–down predictions from the prior/internalized categories. In the simulation of Fig. 4 the relative precision of visual and acoustic cues were chosen so that both modalities had a similar weight on the inference process, as evidenced by the strong activation of both /aba/ and/ada/ units for combination stimuli, which are characterized by strongly informative acoustic and visual cues (the top right corner of the figure). We assume that this situation leads to a maximum of combination and fusion responses to combination-generating (top right region) and fusion-generating (bottom left region) stimuli, respectively.

When the activity of one unit predominates over the others, as is the case in the presence of strong evidence for one of the three speech tokens, it is effectively making a prediction about the identity of the stimulus. If the evidence comes primarily from one of the two modalities (because of its salience or timing) the top–down generative part of the model predicts the sensory input on the other modality. If the actual sensory input in the other modality is sufficiently close to the top–-down predicted input, it reinforces the current estimate. This is the case for congruent /aba/ or /aga/. The opposite is observed for combination stimuli, for which both acoustic /aga/ and visual /aba/ provide strong evidence (for /aga/ and /aba/ respectively). The prediction from one modality is not confirmed in the other modality and the model cannot provide a single consonant solution. The activity in the top right panel of Fig. 4 shows that both the /aba/ recognition unit (black) and the /aga/ recognition unit (dashed) are strongly activated at different times of the stimulus presentation.

Starting with the parameters used for the simulations shown in Fig. 4, we varied the relative precision of visual and acoustic streams at the lower level of the model and run the recognition task on (12 lip closure × 13 F2) pairs. The central panel in Fig. 5 corresponds to the same parameters as those in Fig. 4. The recognition unit most active at the end of the process is now represented by the color of the filled circle. Some plots in Fig. 4 (diagonal region in the upper right quadrant) show strong activation of /aba/ and /aga/ recognition units alternating in time. This alternation reflects cross-modal conflict and is represented in Fig. 5 by the size of the filled circles, with smaller circles representing less congruence, i.e., more conflict. As an example we may compare the size of the circle for the standard combination in Fig. 5B, marked by the black circle on the upper right quadrant, and the corresponding recognition unit activations for the same stimulus in Fig. 4 (gray dashed square). Small size circles denote that the dominant recognition unit varies (alternates), signaling conflicting stimuli. Larger circles in Fig. 5B, indicate a steady ramp up of a single recognition unit when stimuli are not conflicting.

When the strength of the visual cue was strengthened by decreasing the amplitude of fluctuations corresponding to x_L (the lip aperture variable) and increasing the amplitude of the fluctuations for x_F (the second formant variable), the pattern of activity in the recognition units changed in the stimulus region around the largest cross-modal conflict (upper right quadrant). There was a shift from dominant /aga/ (light grey) to dominant /aba/ (black) and the conflict region (smaller sized circles) shifted downwards in the direction of /aga/ (Fig. 5A). That is, some stimuli that were perceived as /aga/ (control, Fig. 5B) changed to /aba/ (Fig. 5A) and for other stimuli conflict increased (reduction of circle size). The same qualitative changes, but in the opposite direction (from /aba/ to /aga/) were obtained when the amplitude of the fluctuations for x_F was decreased and those for x_L increased, leading to a relative strengthening of the acoustic cue (Fig. 5C). The winner-take-all dynamics in the model also suggests that earlier arriving cues could have an advantage over later cues. We found a subtle effect consistent with such an advantage (Fig. A.2). When the model encoded AV patterns with later arriving acoustic cues (Appendix 1, Fig. A.2A), the /aba/ recognition unit in the cross-modal conflict region (stimuli close to a visual /aba/ and auditory /aga/) was favored, suggesting a relative advantage of the visual cue. With patterns with earlier acoustic cues (Fig. A.2C), it was the activity of the /aga/ recognition unit that showed increased activity and lower conflict in the visual /aba/, auditory /aga/ region of stimulus space, thus pointing to a relative advantage of the acoustic cues. The other regions of stimulus space, which are closer to congruent and fusion stimuli, underwent only modest changes (Appendix 1).

In summary, since the recognition units at the top level are multimodal and send top–down connections to the individual sensory modalities, early salient information about the stimulus in one sensory modality is transferred to the other modality in the form of predictions of sensory input (through the top–down generative component of the model). Our simulations suggest that fusion occurs because cross-modal predictions are consistent, and the brain does not detect incongruence. Conversely, combination percepts occur because cross-modal predictions are violated.

4. Discussion

To investigate the role of predictions across modalities in the early stages of audiovisual speech processing, we explored

how a hierarchical predictive coding model with acoustic and visual sensory input would process incongruent audiovisual speech tokens. The model was designed to produce synthetic versions of real congruent AV productions of /aba/, /ada/ and /aga/from a single speaker. The model's performance was subsequently tested on arbitrary incongruent combinations of acoustic and visual cues.

Optimal probabilistic inference provides a principled and useful framework to describe basic features of multisensory cue integration (Pouget, Beck, Ma, & Latham, 2013) such as reliability-based cue weighting and the brain's interpretation of sensory stimuli based on prior knowledge. Predictive coding, in addition, considers the brain as a predictive machine, with higher cortical areas trying to explain/predict activity in lower areas (Clark, 2013; Rao & Ballard, 1999). It naturally incorporates top–down information and can deal with time varying continuous inputs. In this work we used a predictive coding framework based on Dynamic Expectation Maximization (Friston et al., 2008) that has previously been used to model acoustic speech processing (Yildiz et al., 2013). Yildiz et al. modeled words or sentences individually with a single module encoding a 6 band acoustic cochleogram of a spoken word or sentence. Our predictive coding model included the visual modality in addition to the acoustic modality and we explicitly included several patterns (/aba/, /ada/, /aga/) competing to explain the incoming input.

The generative model is a crucial element in predictive coding; it approximates the statistical relationship between an external cause and its sensory consequences, in the present case between the identity of the audiovisual token, and the acoustic and visual sensory cues. We represented the identity of the audiovisual speech tokens with supramodal recognition units at the top level of the hierarchy. There were three recognition units, representing three possible congruent AV speech tokens of the form /aCa/ with C one of the consonants /b/, /d/ or /g/. The top–down generation process of the model drove activity in the sensory units by a weighted sum of memorized/internalized patterns; how much each pattern contributed to the sensory units was determined by the normalized activity of the supramodal recognition units at the top. Such normalization is a candidate basic operation in brain computations (Carandini & Heeger, 2012; Ohshiro, Angelaki, & DeAngelis, 2011).

At the other end of the hierarchical model, there were two sensory units, one per modality. We chose lip aperture as the visual cue and the second formant transitions of the sound spectrogram as the acoustic cue. Although there are other available cues related to the AV speech tokens, lip aperture and second formant were sufficient to distinguish between the congruent productions in our data set. Formant transitions are one of the important acoustic features for plosive consonant classification (Soli & Arabie, 1979; Edwards, 1981; Stevens, 2002) and the second formant transitions were determinant for distinguishing between /b/ and /d/ in noise (Varnet, Knoblauch, Meunier, & Hoen, 2013). The visual contribution to speech is not well understood but it includes both identity and timing information (Brancazio, 2014; D'Ausilio, Bartoli, Maffongelli, Berry, & Fadiga, 2014; Ten Oever, Sack, Wheat, Bien, & van Atteveldt, 2013). Lip aperture is related to articulatory gestures (Browman & Goldstein, 1992) and provides place of articulation information that complements manner of articulation, which is only available acoustically. The internalized lip and second formant patterns associated with the three phonemic segments /b/, /d/, /g/ within the context set by the initial and final vowel /a/ consisted of appropriate target values in the two sensory modalities sampled at approximately 40 Hz.

Amongst the parameters of the model, there were two time scales of special significance, the detail in the pattern representation and the integration time constant at the recognition level. The temporal detail in the representation corresponded to ~20msec (gamma range). The integration time constant, i.e., the time constant of evidence accumulation, was ~200 msec, in the theta range. The integration time constant corresponds to the temporal window of integration, the duration over which features are integrated into a single speech token. This window has been estimated to be about 200 msec in audiovisual speech (Massaro & Cohen, 1993; Van Wassenhove et al., 2007). However, even within the acoustic domain, consonants are characterized by a collection of acoustic features spanning a typical duration of the order of 100 msec (Luo & Poeppel, 2012). This reflects the hierarchy of time scales present in the acoustic speech signal; which contains both fast events ~20 msec such as the onset and offset of vocalic voicing and the broadband burst after the release of the oral cavity occlusion, and slower ~100 msec modulations in the envelope of the speech sound and smooth formant transitions. The online continuous processing presented here opens up the possibility for exploring different temporal scales either nested (Ghitza, 2011) or in parallel.

Precision is also an important parameter of the model, as exemplified by the observation that changing the relative precision of acoustic and visual cues biases the outcomes toward the visual or the acoustic token in combination stimuli (Fig. 5). If precision values, which could be related to pyramidal cell excitability (Brown & Friston, 2012), varied across individuals and attentional state, they would be consistent with the variability of fusion and combination responses across individuals (Nath & Beauchamp, 2012) and under differing attentional settings (Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Tiippana, Andersen, & Sams, 2004).

According to the classification by Schwartz, Robert-Ribes, and Escudier (1998) the model presented here would fall into the direct identification bimodal category; the recognition/classification units being driven/driving audiovisual sensory features without an intermediate unimodal classification stage. Our model is agnostic about the nature of the top-level amodal representations, which here are just abstract symbols (Nearey, 1992). They are encoded as the temporal patterns of sensory activity associated with each of the three phonemic segments /bdg/ within the context provided by the initial and final vowel /a/.

The fuzzy logical model of speech perception (FLMP) is a reference model in audiovisual speech perception (Massaro, 1998) and is related to cue combination based on Bayes theorem. It assumes that the two sensory modalities are processed independently, each providing evidence for each speech token. It has also been applied to fit McGurk perception data. Although it can account for fusion percepts, it cannot account for combination percepts because predictions for the

audiovisual input result from a product of probabilities of responses for unimodal stimuli and combination percepts are very rare for unimodal stimuli. The present model can be viewed as a generalization into an online dynamic inference process. Unlike FLMP, it can detect combinations, under the assumption that the cross-modal conflict identified by the model is occasionally resolved by combination percepts. Another difference is our introduction of top–down predictive signals that bias activity at the unimodal level towards patterns compatible with existing representations at the top level. Therefore, our model does not process modalities independently; the processing in one modality influences the processing in the other one, as suggested by electrophysiological data (Arnal, Wyart, & Giraud, 2011; Van Wassenhove et al., 2005, 2007).

The model is built at an abstract level and it does not implement many features of auditory and visual processing. We implemented the minimum amount of features that are liable to illustrate our main hypothesis, namely, that there are continuous cross-modal predictive signals in general audiovisual processing. AV speech processing provides a useful model to study such predictive signals, and incongruent AV speech tokens illustrate cross-modal predictions that are either easily confirmed, as in the case of fusion percepts, or initially violated and requiring further processing, as in the case of combination percepts. We did not model the process by which cross-modal contradiction is resolved by a combination percept, but we believe that auditory echoic and visual iconic memory could play a role; in the absence of a satisfactory syllabic/coarse level match, the brain might resort to a slower matching process involving more temporal details by combining sub-syllabic features, and comparing the predicted audiovisual features to the content of the sensory buffers, which store about 1 sec of detailed acoustic and visual information. We therefore propose that in the case of combination percepts a match to the incongruent audiovisual stimulus is also found. As Massaro and Cohen (1993) point out, combination percepts can only arise if the resulting percept is "reasonably compatible with the physical properties of a cluster articulation". We suggest that this matching requires processing at a more detailed level, which takes longer to be achieved. This prediction still requires experimental validation.

We did not explicitly implement transmission delays either. Transmission delays are assumed to be larger for visual than acoustic information and this can be implicitly implemented in the model by the relative timing of the predicted second formant and lip closure dynamics, which in the model is encoded in the patterns represented in Fig. 2B. As shown in additional simulations, changing the relative timing in the encoded patterns did not change the main observations (Fig. A.2).

4.1. *Across-modality predictions, fusion versus combination*

Since our generative model was used to create congruent audiovisual speech tokens, its predictive coding version properly processed congruent stimuli by design. The model was then tested on incongruent stimuli. Rather than restricting the model input to the two prototypical incongruent pairings based on our synthetic congruent stimuli, taking the lip aperture from /b/, or /g/ and the second formant from /g/ or /b/, respectively, we explored the model behavior on the two dimensional feature space spanned by lip aperture and second formant. We hypothesize that stimuli represented in this two-dimensional space include the natural variability in speech productions across and within speakers. The location of prototypical congruent and incongruent pairings in this space represented a meaningful metric. Incongruent stimuli evoking fusion responses fell relatively close to the congruent /ada/, while combination evoking incongruent stimuli fell far from congruent tokens, confirming what was noted by Massaro and Cohen (1993) and quantified by Omata and Mogi (2008). This resulted in a qualitative difference in the nature of the predictions/expectations generated at the sensory level by the recognition units. In fusion-generating stimuli the /ada/ token was activated and the ensuing predictions were not far from the actual sensory inputs; sensory predictions were hence confirmed. This contrasted with the combination evoking stimuli. Since both acoustic /aga/ and visual /aba/ provided salient and conflicting information, the expectations on the complementary modality were contradicted, and in our simulations the recognition units representing /aga/ and /aba/ were both strongly activated, though at different times. The activation sequence was determined by the modality dependent timing of consonant specific information, maximal in the intervocalic period for the visual modality and maximal just before and after the intervocalic period for the acoustic modality. Jesse and Massaro (2010) also found differences in the timing of available information from the visual and acoustic modalities in English consonant-vowel-consonant words. Our model also reproduces their finding that early evidence has a relative advantage over later arriving evidence. In the model this comes from the winner-take-all architecture at the recognition level; the more one of the recognition units dominates over the others, the larger the evidence/prediction error needed to change the output at the recognition level (Fig. A.2).

Neurophysiological data also suggest a qualitative difference between fusion and combination. Brain activity in early auditory cortex in fusion trials in which participants do give fused responses, is very similar to that observed upon congruent token presentations (Kislyuk, Möttönen, & Sams, 2008) with differences appearing elsewhere (Erickson et al., 2014). Arnal et al. (2011) compared congruent and incongruent non-McGurk AV speech tokens, leading only to combination percepts when incongruent, and found different patterns of brain activation. The visual input showed two kinds of effects on the perceptual process: an early non-specific effect on auditory cortex and a token-specific effect related to the informative content of the visual modality about the consonant. That is, a prediction error signal, which correlated with increased gamma activity in sensory areas, was detected when the acoustic token invalidated the visual prediction. The observed brain activity changes presumably reflected intrinsically distinct processing for incongruent stimuli that can or cannot be fused.

4.2. Timing and oscillations

In this modeling work, we focused on speech identity information carried by the visual modality. However, the early visually mediated enhancement of auditory cortex activity does not seem to be content specific (Arnal, Morillon, Kell, & Giraud, 2009; Van Wassenhove et al., 2005). It is possible that early detection of lip motion predicts the timing of auditory events and synchronizes the activity of auditory cortex neurons (Arnal & Giraud, 2012; Lakatos, Chen, O'Connell, Mills, & Schroeder, 2007; Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008). The timing units of the model, which here were initialized by hand so that the first sequence unit was in its active state at the beginning of the simulation, could provide an appropriate model for such an early resetting. The dynamic equations at the sequence unit level yield two types of solutions: a periodic mode in which the units get activated in sequence, and an equilibrium mode in which all the units are concurrently activated. We could relate the equilibrium mode with asynchronous activity; a strong input to the unit encoding the beginning of the audiovisual pattern could set the sequence in motion and would be equivalent to resetting/synchronization of oscillations. Future instantiations of the model will be able to address these timing issues.

The sequence unit level of the model was used to extract the right ordering of the stored audiovisual pattern associated with each speech token. How sequences are stored and retrieved by the brain is the subject of ongoing research but sequence representations of the form used here are observed in the firing of the songbird HVC neurons at specific times within a birdsong (Hanhloser et al., 2002). Similarly, sequential activation is also observed in the hippocampus, where it is associated to replay of behavioral experience (Louie & Wilson, 2001) and stored sequence retrieval and replay (Lisman et al., 2009).

4.3. Model extension

The present work should be seen as a proof of principle that predictive coding is a useful framework for addressing the computational bases of speech processing. Although we focused here on the /bdg/ consonant family, lip aperture and F2 modulations in the /ptk/ family are very similar. Distinguishing between /bdg/ and /ptk/ would require augmenting the number of recognition units and the number of sensory features that the model generates, adding a voicing dimension. Extending the model to /ptk/ would also delay the time at which one of the recognition units dominates, since the discriminant cues between /ptk/ and /bdg/ occur later than those needed to distinguish tokens within the /bdg/ family. However, the inclusion of further recognition units would still preserve the qualitative differences between fusion and combination, and we predict that the model would give similar results on all classes of phonemic families. Beyond such a straightforward extension, an additional level of processing able to resolve the conflict occurring in combination conditions by explicitly resorting to phonemic-level recognition units could be added. This would allow more precise predictions about the conflict resolution mechanism and would therefore help interpreting the patterns of brain activity in response to incongruent audiovisual speech tokens (Arnal et al., 2011).

5. Conclusion

In this contribution to this special Cortex issue, we have presented a minimal predictive coding model of incongruent audiovisual speech processing. The model architecture permits the dynamic deployment of predictions across modalities. We found that fusion happens because the representation of fusion-evoking stimuli in the two-dimension lip/F2 space fall close to the representation of a congruent stimulus, hence meeting expectations for a congruent stimulus. Therefore fusion percepts are not accompanied by any mismatching sensation. In contrast, combinations happen because each stimulus modality generates strong conflicting predictions about the other modality.

Acknowledgments

We would like to thank Jean Luc Schwartz for helpful discussions and comments on the manuscript. IO, SB and ALG are funded by the European Research Council (Compuslang project; Grant agreement 260347) and the SNF (The Swiss National Science Foundation, number 320030_149319).

Appendix

Two-dimensional visuo-acoustic stimulus space

Fig. A.1 shows the recognition process shown in Fig. 4 but on a finer scale, which allows a more precise representation of the /ada/ congruent token. It also shows that with our parameters, the /aba/-/ada/ transition is smoother than the /aga/-/ada/ transition. This is because /aba/ and /ada/ differ mostly in the visual dimension, which has a lower precision ($\log P(x_L) = \log P(v_L) = 5$; $\log P(x_F) = \log P(v_F) = 8$)./ada/ differs from /aga/ mostly in the acoustic dimension, which has a higher precision and therefore a more abrupt transition.

Relative cue timing

Fig. A.2 explores the role of relative cue timing across modalities. It shows three hypothetical lip and F2 patterns for congruent /aba/, /ada/ and /aga/ tokens. The patterns differ in the relative timing of peak lip closure and F2 onset and offset times. For each pattern we generated congruent stimuli and then combined them with the same (A_F and A_L) combinations to test performance for a wide range of lip closure and F2 pairings.

The recognition performance with all other parameters unchanged shows a change in the activity of recognition units presented with lip and F2 mixtures close to the diagonal of the upper right quadrant. The recognition unit most activated at the end of stimulus presentation changes from /aga/ to /aba/ when visual information, which is more salient for /aba/, arrives earlier. Conversely, the most activated unit changes from /aba/ to /aga/ when the visual information arrives later (Fig. A.2).

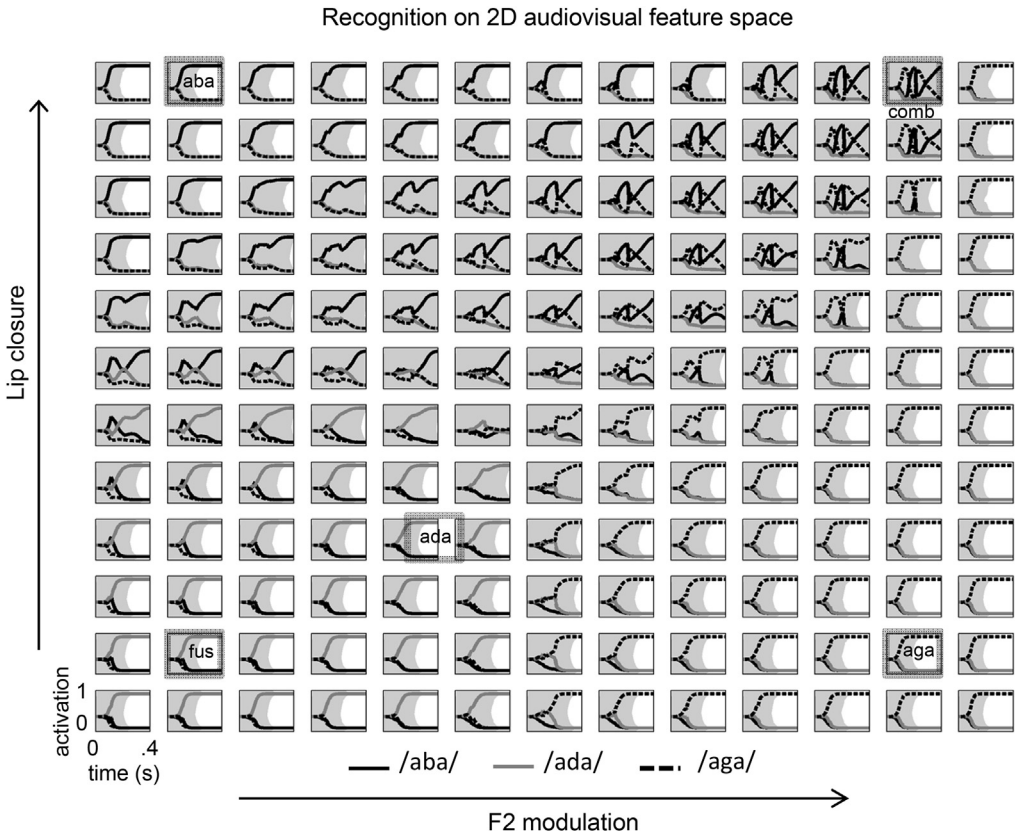

Fig. A.1 – Model performance on a two-dimensional family of lip closure and F2 paired stimuli. A more detailed sampling of the space with the same parameters used in Fig. 4.

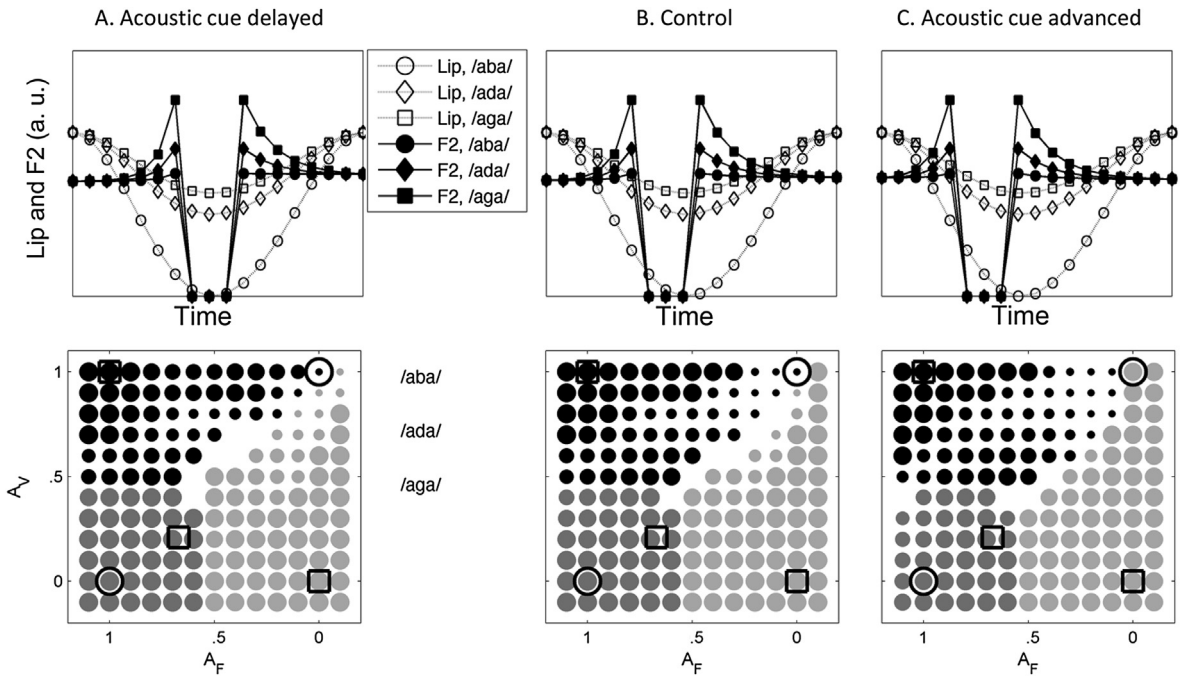

Fig. A.2 – Performance of two additional versions of the model. Encoded patterns differ in the relative timing of acoustic and visual information (top row). The bottom row shows the response to12 × 13 pairs of lip aperture and F2 (see Methods for a detailed description). The position on the grid determines the relative distance from the /aba/ stimulus in the lip closure dimension (A_V) and F2 modulation dimension (A_F). The positions of the standard /aba/ ($A_V = 1$, $A_F = 1$), /ada/ ($A_V = .21$, $A_F = .67$) and /aga/ ($A_V = 0$, $A_F = 0$) are marked with a square, and those of the standard fusion ($A_V = 0$, $A_F = 1$) and combination stimuli ($A_V = 1$, $A_F = 0$) by a circle. At each point of the grid, the color of the filled circles represents the identity of the recognition unit with highest activity at the end of the process. The size of the filled circle is inversely related to the fluctuations in recognition unit activity (smaller circles related to larger cross-modal conflict). As the acoustic cue is progressively advanced with respect to the visual cue (left to right), some of the responses to stimuli in the visual /aba/ and acoustic /aga/ region (close to the black circle in the upper right quadrant) become more like /aga/: 1) /aba/ activations show more conflict (smaller circles) and /aga/ activations show less conflict (larger circles), and 2) some stimuli change from maximally activated /aba/ to maximally activated /aga/.

REFERENCES

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology, 15*, 839–843.

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences, 16*(7), 390–398.

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A.-L. (2009). Dual neural routing of visual facilitation in speech processing. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 29*(43), 13445–13453. http://dx.doi.org/10.1523/JNEUROSCI.3194-09.2009.

Arnal, L. H., Wyart, V., & Giraud, A.-L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience, 14*, 797–801.

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: insights from audio-visual speech perception. *PloS One, 6*(5), e19812.

Bendixen, A., Scharinger, M., Strauß, A., & Obleser, J. (2014). Prediction in the service of comprehension: modulated early brain responses to omitted speech segments. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 53*, 9–26.

Boersma, Paul, & Weenink, David (2014). *Praat: Doing phonetics by computer [Computer program]*. Version 5.4.

Brancazio, L. (2014). Measuring visual contributions in phonetic categorization. *The Journal of the Acoustical Society of America, 135*(4), 2256.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica, 49*(3–4), 155–180.

Brown, H. R., & Friston, K. J. (2012). Dynamic causal modelling of precision and synaptic gain in visual perception - an EEG study. *NeuroImage, 63*(1), 223–231. http://dx.doi.org/10.1016/j.neuroimage.2012.06.044.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience, 13*(1), 51–62.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLOS Computational Biology, 5*, e1000436.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences, 36*(3), 181–204.

D'Ausilio, A., Bartoli, E., Maffongelli, L., Berry, J. J., & Fadiga, L. (2014). Vision of tongue movements bias auditory speech perception. *Neuropsychologia, 63*, 85–91.

Edwards, T. J. (1981). Multiple features analysis of intervocalic English plosives. *The Journal of the Acoustical Society of America, 69*(2), 535–547.

Erickson, L. C., Zielinski, B. A., Zielinski, J. E. V., Liu, G., Turkeltaub, P. E., Leaver, A. M., et al. (2014). Distinct cortical

locations for integration of audiovisual speech and the McGurk effect. *Frontiers in Psychology*, 5, 534.
Friston, K. J., Trujillo-Barreto, N., & Daunizeau, J. (2008). DEM: a variational treatment of dynamic systems. *Neuroimage*, 41, 849–885.
Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology: CB*, 22(7), 615–621.
Ghitza, O. (2011). Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm (June) *Frontiers in Psychology*, 2, 130.
Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, 11, 473–490.
Hahnloser, R. H. R., Kozhevnikov, A. A., & Fee, M. S. (2002). An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature*, 419, 65–70.
Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception & Psychophysics*, 72(1), 209–225.
Kislyuk, D. S., Möttönen, R., & Sams, M. (2008). Visual processing affects the neural basis of auditory discrimination. *Journal of Cognitive Neuroscience*, 20(12), 2175–2184.
Lakatos, P., Chen, C.-M., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron*, 53(2), 279–292.
Lisman, J. E., & Redish, A. D. (2009). Prediction, sequences and the hippocampus. *Philosophical Transactions of the Royal Society of London B Biological Sciences*, 364, 1193–1201.
Louie, K., & Wilson, M. A. (2001). Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, 29, 145–156.
Luo, H., & Poeppel, D. (2012). Cortical oscillations in auditory perception and speech: evidence for two temporal windows in human auditory cortex (May) *Frontiers in Psychology*, 3, 170.
Magnotti, J. F., & Beauchamp, M. S. (2014). The noisy encoding of disparity model of the McGurk effect. *Psychonomic Bulletin & Review*.
Magnotti, J. F., Ma, W. J., & Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech. *Frontiers in Psychology*, 4, 798.
Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to behavioral principle*. Cambridge, Mass: MIT Press.
Massaro, D. W., & Cohen, M. M. (1993). *Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables*.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–748.
Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, 58(3), 351–362.
Nath, A. R., & Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *NeuroImage*, 59(1), 781–787.
Nearey, T. M. (1992). Context effects in a double-weak theory of speech perception. *Language and Speech*, 35(1–2), 153–171.
Ohshiro, T., Angelaki, D. E., & DeAngelis, G. C. (2011). A normalization model of multisensory integration. *Nature Neuroscience*, 14(6), 775–782.
Omata, K., & Mogi, K. (2008). Fusion and combination in audio-visual integration. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 464(2090), 319–340.
Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology*, 3, 320.
Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, 16(9), 1170–1178.
Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87. http://dx.doi.org/10.1038/4580.
Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences*, 12(3), 106–113. http://dx.doi.org/10.1016/j.tics.2008.01.002.
Schwartz, J.-L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: a taxonomy of models for audio- visual fusion in speech perception. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychlogy of speechreading and auditory-visual speech* (pp. 85–106). Hove, UK: Taylor & Francis.
Schwartz, J.-L., & Savariaux, C. (2014). No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLoS Computational Biology*, 10(7), e1003743.
Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *Journal of Neuroscience*, 32, 8443–8453. c.
Soli, S. D., & Arabie, P. (1979). Auditory versus phonetic accounts of observed confusions between consonant phonemes. *The Journal of the Acoustical Society of America*, 66(1), 46–59.
Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 111(4), 1872. http://dx.doi.org/10.1121/1.1458026.
Stevenson, R. A., & Wallace, M. T. (2013). Multisensory temporal integration: task and stimulus dependencies. *Experimental Brain Research*, 227, 249–261.
Summerfield, Q. (1987). Some preliminaries to a comprehensive account\rof audiovisual speech perception. In B. Dodd, & R. Campbell (Eds.), *Hearing by eye: The psychology of lipreading* (pp. 3–51). Lawrence Erlbaum.
Ten Oever, S., Sack, A. T., Wheat, K. L., Bien, N., & van Atteveldt, N. (2013). Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs. *Frontiers in Psychology*, 4, 331.
Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology*, 16, 457–472.
Van Wassenhove, V. (2013). Speech through ears and eyes: interfacing the senses with the supramodal brain. *Frontiers in Psychology*, 4, 388.
Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 1181–1186.
Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, 45, 598–607.
Varnet, L., Knoblauch, K., Meunier, F., & Hoen, M. (2013). Using auditory classification images for the identification of fine acoustic cues used in speech perception. *Frontiers in Human Neuroscience*, 7, 865.
Yildiz, I. B., von Kriegstein, K., & Kiebel, S. J. (2013). From birdsong to human speech recognition: bayesian inference on a hierarchy of nonlinear dynamical systems. *PLOS Computational Biology*, 9, e1003219.