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/© 2015 Published by Elsevier Ltd.

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