The McGurk effect in symmetric neural networks*

Klaas T. Seinhorst
Department of Psychology, University of Amsterdam

This article brings together two topics on which Paul Boersma and I have worked together: the interaction of auditory and visual cues in speech perception, and symmetric neural networks as a decision mechanism. The McGurk effect, first described by McGurk and MacDonald in 1976, reveals how seer-listeners integrate conflicting visual and auditory information, and how the former can even override the latter. Boersma provided an Optimality-Theoretic formalization of this process; in this article, I present two neural network models to show that the McGurk effect is an emergent epiphenomenon in the perception of speech.

1 Introduction (1): integrating visual and auditory information

In 2009, Paul supervised my bachelor's thesis (Seinhorst 2009), in which I described the adaptation of the high front rounded vowel [y] into Russian. Russian generally borrows this vowel as /u/ with palatalization of the preceding consonant (for instance, French *étude* [etyd] 'etude' has been borrowed as /et^{j}ud/, and German *Düne* [dynə] 'dune' as /d^{j}una/),[1] unless that consonant resists palatalization or is lacking altogether (i.e., word-initially). Using Optimality Theory (Prince & Smolensky 1993/2004), I first attempted a formalization combining only auditory cues and phonotactic restrictions, as Boersma and Hamann (2009) had done to explain other instances of loanword adaptation. However, the relevant phonetic data on French, German and Russian F_2 and F_3 often suggested that Russian /$^{(j)}$i/ was the best match for foreign [y], while the data clearly suggested /$^{(j)}$u/. I ended up incorporating visual information into my formalization, assuming that the clearly visible lip protrusion in [y], similar to that in [u], overrides the difference in F_2 and steers the seer-listener to /u/. This approach built upon earlier work by Paul, who had already described the McGurk

**** Thanks to the guest editors and reviewer, and to Paul for all the input and inspiration he has provided over the years. Er zijn weinig mensen die drie diploma's bij dezelfde begeleider behalen: het feit dat ik dat wél heb gedaan bij jou, laat zien hoezeer onze interesses overlapten, en hoeveel ik van je kon en wilde leren.

[1] I will explain the different notations when I describe the theoretical model in 1.2.

© Klaas Seinhorst

effect (McGurk & MacDonald 1976) as an interaction of auditory, visual and phonotactic considerations (Boersma 2006/2012).

1.1 The McGurk effect: McGurk & MacDonald (1976)

Almost fifty years before this article appeared, psychologists Harry McGurk and John MacDonald published results from experiments that show how visual cues can override, and be integrated with, auditory cues in speech perception. On video tapes, they synthesized repetitions of CV syllables consisting of a plosive [b d g] plus [a] sequence in which the onsets in the auditory modality did not match the onsets in the visual modality, so for instance auditory [baba] with visual [gaga]; they then presented participants with the stimuli in two conditions, one in which participants saw and heard the speaker, and one in which they only heard the speaker. Participants were asked to repeat what they perceived.

In the auditory-only condition, accuracy was high; however, in the auditory-visual condition, participants perceived the combination of auditory [baba] and visual [gaga] as /dada/, and the combination of auditory [gaga] and visual [baba] as a percept that contained both /b/ and /g/, such as /gabga/ or /baga/. McGurk and MacDonald state that in the absence of auditory input, people confuse alveolars and velars with each other but never with labials, and they never confuse labials with alveolars or velars.

McGurk and MacDonald hypothesized that [baba] and [dada] share some auditory information, or else auditory [baba] coupled with visual [gaga] could not be perceived as the "unifying percept" /dada/; they do, however, not comment on what exactly this auditory similarity might be.

Since 1976, the effect has been studied extensively. It appears to be present already in 4.5-month old infants (Burnham & Dodd 2004); it occurs in vowels too (Traunmüller & Öhrström 2007; Valkenier et al. 2012); and some cultures are apparently more sensitive to it than others, with Japanese seer-listeners being less susceptible to the effect than American English seer-listeners (Sekiyama & Tohkura 1991; Sekiyama & Burnham 2008), although the Japanese listeners of Sekiyama and Tohkura did rely more on the visual modality if noise was added to the auditory stimuli; in fact, for people with hearing impairment, visual information is often decisive in the perception of speech (Miller & D'Esposito 2005; Başkent & Bazo 2011).

1.2 The McGurk effect: Boersma (2006/2012)

Boersma (2006/2012) formalized the McGurk effect within his theoretical model of bidirectional phonology and phonetics (BiPhon; see Boersma 2011 for a comprehensive overview and references of the proposed representations and mappings), which assumes that the speaker–listener uses the same knowledge in speech production (i.e., as a speaker) and perception (i.e., as a listener). Figure 1

shows the phonological and phonetic representations in the model, as well as the ways in which they are connected.

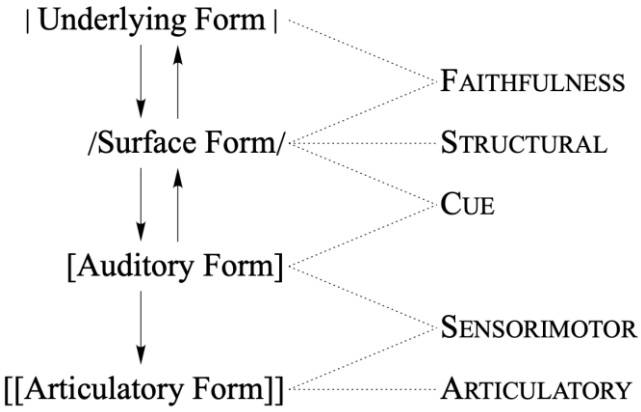


Figure 1: The BiPhon model of speech perception and production. A speaker travels downwards through the figure, from an Underlying Form to an Articulatory Form; a listener travels upwards, from an Auditory Form to an Underlying Form.

The Underlying Form (UF) is a discrete lexical representation, structured in terms of morphemes and phonemes; the Surface Form (SF) is a discrete prosodic representation, consisting of prosodic units such as syllables; an Auditory Form (AudF) contains continuous auditory information such as formants and frication noise; and an Articulatory Form (ArtF) specifies the muscle movements that are needed to pronounce an utterance.

Faithfulness knowledge tries to ensure that no mismatches occur in the mapping between UF and SF (e.g. "don't insert a segment"); structural knowledge tries to ensure the phonotactic well-formedness of the SF (e.g. "do not start a syllable with a complex onset", if the language has such a restriction); cue knowledge links phonological, discrete units to continuous auditory information and vice versa (e.g. "a plosive release burst does not signal a vowel segment"); sensorimotor knowledge evaluates the relationship between auditory information and the muscle movements needed to produce it; and articulatory knowledge allows a speaker to assess which movements require more effort than others.

A speaker starts from an Underlying Form and ends up pronouncing an Articulatory Form, while a listener starts from an Auditory Form and ends up recognizing an Underlying Form. All knowledge is used bidirectionally: for instance, the same knowledge that prevents a speaker–listener from pronouncing a nasal with a plosive release burst prevents them from perceiving a plosive release burst as a nasal. Representations and connections are evaluated in parallel and not serially, which entails that a "later" process may influence an "earlier"

one: for instance, in production, articulatory considerations may override phonological ones.

Boersma formalizes the McGurk effect in Optimality Theory, as a mapping of continuous auditory and visual cues onto a discrete phonological representation: in Figure 1 this would be the arrow from a mixed auditory and visual form (a "Sensory Form") to the Surface Form. This process involves three types of constraints: auditory cue constraints, visual cue constraints, and structural constraints. I will first introduce the necessary cue constraints.

The auditory property that [b] and [d] share with each other but not with [g], which McGurk and MacDonald did not specify, is that their F_2 and F_3 are not close together, while [g] displays the so-called "velar pinch". This yields an auditory cue constraint "*/g/ [separated F_2 & F_3]" '/g/ is not signalled by a separated F_2 and F_3', which must be high-ranked. (Boersma foregoes the constraints linking separated F_2 and F_3 to /b d/, as they will be low-ranked and not decisive.)

In Boersma's formalization, the auditory property that distinguishes /b/ and /d/ is their F_2, with the latter having a higher F_2 than the former. Boersma includes all three auditory cue constraints "*/b/ [low F_2]", "*/d/ [low F_2]" and "*/g/ [low F_2]", in perception meaning: "do not perceive a low F_2 as /*x*/".

In the visual modality, the lip closure seen in [b] does not map to /d/ or /g/, or conversely, the open lips seen in [d] and [g] do not map to /b/, yielding the three visual cue constraints "*/b/ [open lips]", "*/d/ [open lips]" and "*/g/ [open lips]", only the first of which must be high-ranked.

Boersma combines these seven constraints into the following tableau of the perception of auditory [ba] and visual [ga]:

Tableau 1: Perceiving visual [ga] and auditory [ba] (Tableau (1) from Boersma 2006/2012: 303).

[open lips, separated F_2&F_3, low F_2]	*/b/ [open lips]	*/g/ [separated F_2 & F_3]	*/d/ [low F_2]	*/g/ [low F_2]	*/b/ [low F_2]	*/d/ [open lips]	*/g/ [open lips]
/ba/	*!				*		
☞ /da/			*			*	
/ga/		*!		*			*

This tableau neatly shows how the "unifying percept" /da/ arises. The visual cue overrules all auditory cues and immediately rules out /ba/ as a candidate, thus leaving /da/ and /ga/; because F_2 and F_3 in the input are separated, /ga/ is ruled out too, leaving /da/ as the winner. (Although Boersma separates them with a solid line, the mutual ranking of "*/b/ [open lips]" and "*/g/ [separated F_2 & F_3]" cannot

be established, so it is unclear whether the perception proceeds in this stepwise manner, but the result remains the same if it does not.)

Structural constraints are necessary to explain the reverse McGurk effect, the combination of auditory [g] and visual [b], where no such "unifying percept" is available and seer-listeners maintain both [g] and [b]. Because English does not allow for syllables that start with a labial plus velar sequence, nor for syllables without vowels, participants distribute the plosives to different syllables, mapping the input $[gaga]_{Aud}$ plus $[baba]_{Vis}$ to the attested "combination" percepts such as /bag.ba/. This is shown in Boersma's Tableau (4), repeated here as Tableau 2. The cue constraints are formulated positively here, rather than negatively as in Tableau 1: in perception, "[closed lips] ⇒ /lab/" means 'if you see closed lips you *must* perceive a labial'.

Tableau 2: Perceiving visual [ba] and auditory [ga] (Tableau (4) from Boersma 2006/2012: 306)

[closed lips, close F_2&F_3]	*/.labvel/	*/.C./	[closed lips] ⇒ /lab/	[close F_2&F_3] ⇒ /vel/
/.ga.ga./			*!	
☞ /.bag.ba./				*
/.gbag.ba./	*!			
/.gba.gba./	*!*			
/.g.bag.ba./		*!		
/.ba.ba./				*!*

The tableau shows that although a candidate like /.gba.gba./ preserves the auditory and visual cues, it is not a viable winner in English because of the top-ranked structural constraint against labial plus velar sequences.

McGurk and MacDonald report that some participants repeat "baga"; while this candidate violates both cue constraints once, it does obey the structural constraints. Its occurrence is unexpected in Tableau 2, which would require an additional cue constraint that requires both [closed lips] and [close F_2&F_3] to have an equivalent in the output, assigning violations to candidates that lack a labial and a velar; such a constraint would explain why /.ba.ga./ is a better candidate than /.ga.ga./ or /.ba.ba./.

2 Introduction (2): symmetric neural networks

In 2012, Paul supervised my master's thesis (Seinhorst 2012), in which I modelled the emergence of auditory contrast with symmetric neural networks as a decision mechanism. As Boersma and Hamann (2008) had done with OT, I showed that

optimal auditory dispersion arises because speakers learn to prefer unambiguous auditory cues in perception, but resist investing more articulatory effort than necessary when they use their cue knowledge bidirectionally: when these two effects balance each other out, a stable state with optimal contrast emerges.

Neural networks are models that are composed of nodes, which may be organized in layers, and which may be linked by connections. These connections can be excitatory or inhibitory. Activity can flow through the network, activating or deactivating nodes through the connections. The advantage of neural networks over Optimality Theory is that they can be used to model emergent and gradient phenomena, such as phonological category creation and the perceptual magnet effect (Boersma, Benders & Seinhorst 2020), whereas in OT categories are pre-existing (i.e., already known) and categoryhood is discrete (i.e., yes or no). The neural networks in this article are symmetric because two nodes influence each other to the same degree in both directions of processing, which means that the network can be used to model speech production as well as speech perception, warranting the bidirectionality of the BiPhon model. In the sections below, I will touch upon those aspects of the model that are relevant for my simulations; for an extensive description of its workings, please refer to Boersma, Benders & Seinhorst (2020).

3 The first neural network model of the McGurk effect (two levels, discrete phonetics)

I will first provide the simplest model possible. It consists of two levels, both corresponding to a level of representation: phonology and phonetics (so probably the surface form and sensory form, respectively, as in Boersma 2006/2012). In the phonetics, each relevant cue is represented by a group of four nodes: I use the cues $[\text{low } F_2]_A$, $[F_2 \sim F_3]_A$ (meaning F_2 and F_3 are close to each other), $[\text{closed}]_V$ and $[\text{open}]_V$. In the phonology, the response categories /b/, /d/ and /g/ are each represented by a group of four nodes. In the initial state of the network, all nodes have small, random activities, and all cues are connected to all phonological categories by excitatory cue connections with a small, random weight; at this point, none of the phonological categories is associated with any of the phonetic cues in particular. Every node of a phonological category is connected to every node of the other two phonological categories by an inhibitory connection with a fixed weight of -0.12, implementing the notion of competitive learning (Grossberg 1976, 1987; Rumelhart & Zipser 1985). These connections turn out to play a crucial role in the perception process, as we will see in 3.3. The maximum activity of an unclamped node, a node whose activity is determined by activation and inhibition by adjacent nodes rather than by my determining it and keeping it fixed, is 2.

The model is shown in its initial state in Figure 2. Each circle represents a node, and each line represents a connection. Excitatory connections are drawn in black, inhibitory connections in grey. The stronger a connection is, the thicker it is drawn; the more active a node is, the more filled it is drawn. Nodes drawn with solid lines are clamped, nodes drawn with dotted lines are not.

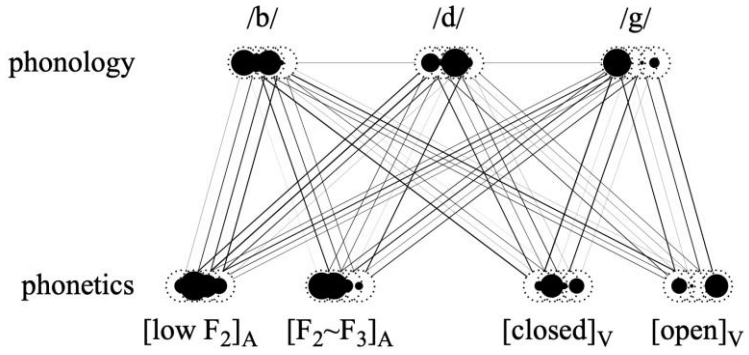


Figure 2: The two-level model in its initial state. Sensory cues are shown on the phonetic layer, response categories on the phonological layer.

This model is closest to Boersma's OT formalization in two ways: the phonological categories are not emergent but given (pre-existing), and the sensory cues are discrete rather than continuous.

3.1 The learning process

In this model, learning means to associate sensory cues with phonological categories. This process happens as follows. In every learning step, a phonological category is drawn at random; its nodes at the phonological layer are then given an activity of 1, and the nodes of the other categories are given an activity of 0. At the phonetic layer, the corresponding cues are activated while the remaining cues are deactivated. All nodes are then clamped, and the weights of the cue constraints are updated according to the "inoutstar" rule (Boersma, Benders & Seinhorst 2020).

In the phonetics, I want to make sure that /b/ and /d/ are similar auditorily while standing out from /g/; to achieve this, I activate the $[F_2{\sim}F_3]_A$ nodes for /g/ but not for /b d/, and I activate the [low F_2]$_A$ nodes for /b d/ but not /g/, although a bit more for /b/ than for /d/, because the former has a lower F_2 than the latter (see Boersma's constraint ranking in Tableau 1). Visually, /d/ and /g/ must be identical to each other but different from /b/. In these simulations, the possible activities at the phonetic layer are as follows:

Table 1: Possible activities at the phonetic layer.

category	$[\text{low } F_2]_A$	$[F_2 \sim F_3]_A$	$[\text{closed}]_V$	$[\text{open}]_V$
/b/	1	0	1	0
/d/	0.5	0	0	1
/g/	0	1	0	1

Figure 3 represents a possible first learning step; the model learns that the phonological category /g/ is associated with an auditory F_2 and F_3 that are close together, and with visually open lips.

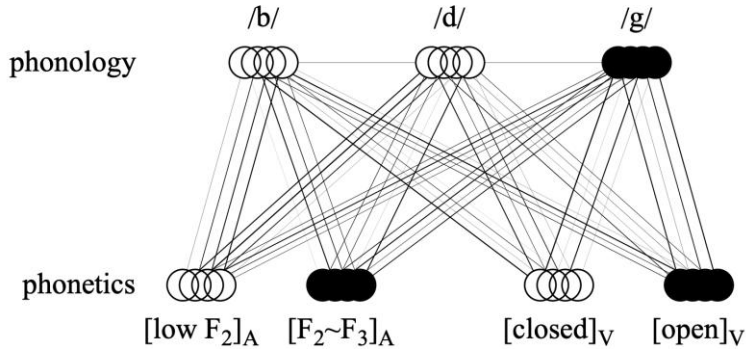


Figure 3: A learning step in the two-level model.

After some 2,000 learning steps, the weights in the network do not change anymore: it has reached a mature state, cf. Figure 4.

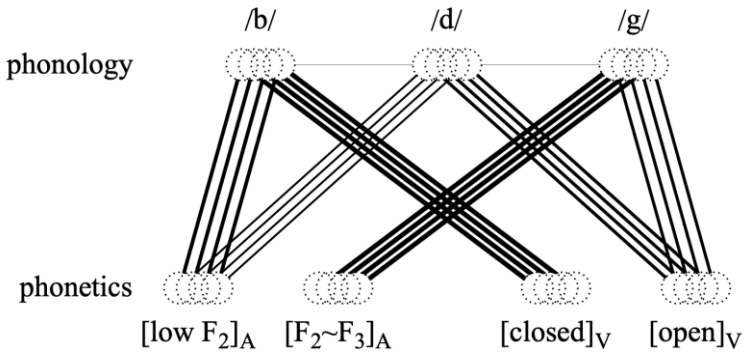


Figure 4: The mature two-level model.

We see that the phonological categories have now clearly become associated with certain phonetic cues, but no longer with others.

3.2 Production and perception

The network was designed to be used bidirectionally, so it can both listen and speak. For instance, I can ask it to produce the /d/ category, which means that it tells us which phonetic cues belong to that category. Doing so entails a few steps:

I set the activities of the /d/ nodes to 1 and the activities of the other phonological nodes to 0; I set the activities of all nodes in the phonetic layer to 0; I clamp all nodes at the phonological layer and unclamp those at the phonetic layer; I let activity spread through the network. Figure 5 shows the result of these steps.

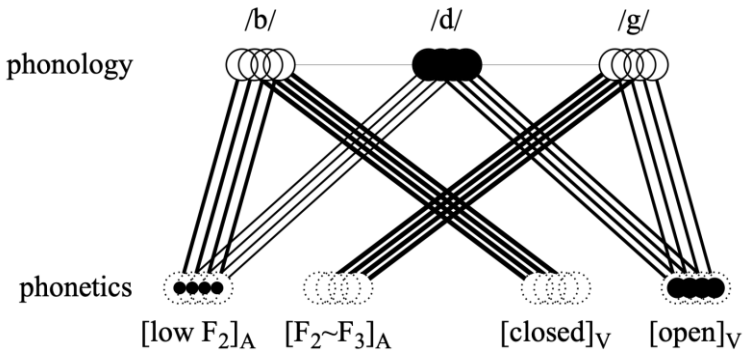


Figure 5: The two-level model producing /d/.

We see that the [low F_2] nodes are somewhat activated, but judging by the weights of the connections less than they would be for /b/, which reflects the difference in their activations in the learning process. We also see that the network produces /d/ with open lips, as expected.

Perceiving works the opposite way: I activate the nodes of the relevant cues and deactivate the other nodes at the phonetic layer, I deactivate all nodes at the phonological layer, and I let activity spread through the network gradually (i.e., in a number of steps) while keeping only the phonetic nodes clamped. This will activate nodes at the phonological layer, and eventually only one category will remain active. This will be the output of the perception process.

Depending on the cue, the activity spreading may activate multiple categories: for instance, because $[open]_V$ is connected to /d/ as well as /g/, activating $[open]_V$ will activate both categories. However, because of the inhibitory lateral connections, as a category becomes more active it will also deactivate the other categories. I will discuss this process in more detail in the following subsection.

3.3 Does the McGurk effect happen?

Of course, we want to test whether the network mimics the McGurk effect when we feed it incongruent cues. In Figure 6 I have activated the $[low\ F_2]_A$ nodes and deactivated the $[F_2{\sim}F_3]_A$ nodes, the pattern that belongs to auditory [b], and I have deactivated the $[closed]_V$ nodes and activated the $[open]_V$ nodes, the pattern that belongs to visual [g].

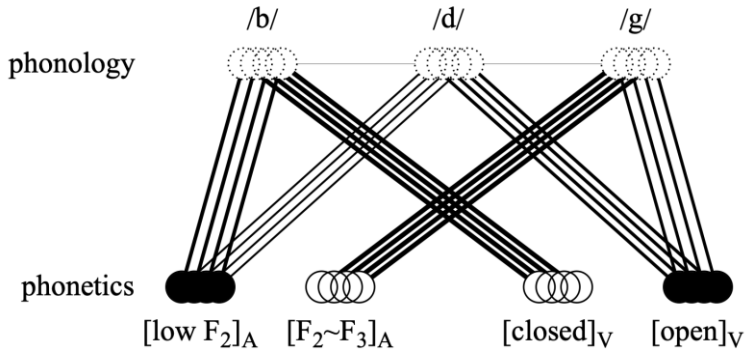


Figure 6: The network perceiving auditory [ba] and visual [ga], before activation spreading.

The two-level model is actually a clarifying visualization, already allowing us to see that the activity from [low F_2]$_A$ will spread both to /b/ and, to a lesser extent, to /d/; the activity from [open]$_V$ will spread to both /d/ and /g/ equally. Although this topography already suggests that /d/ is the "unifying percept" in this case, we will need to confirm that /d/ becomes the most activated, while all three phonological categories will become active. How is this competition settled?

Figure 6 illustrates that /d/ will become activated through two paths, while /b/ and /g/ only become activated through a single path. The sum of the weights of all eight excitatory connections to /d/ is 4.54; the weights of the eight connections to /b/ add up to 3.16, those to /g/ to 2.43. As activity spreads through the network, /d/ will therefore receive more activity than its competitors; consequently, it will start deactivating its rival categories /b/ and /g/ through the lateral inhibitory connections more strongly than vice versa, as a result of which the activities of /b g/ eventually drop to zero. Figure 7 shows the activities of the nodes at the phonological layer with every step of activation spreading (remember that the maximum activity of unclamped nodes was capped at 2):

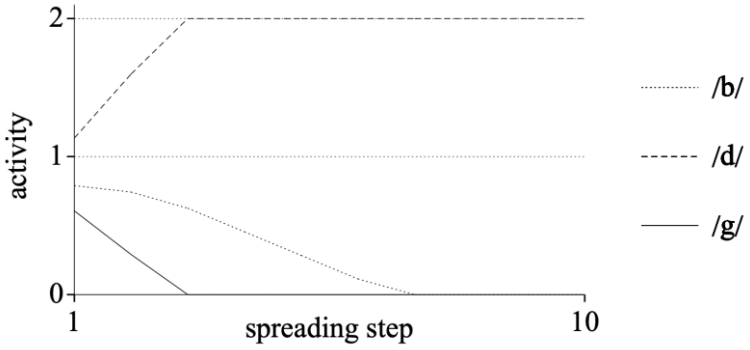


Figure 7: The activities at the phonological layer as activity spreads through the network.

We observe that while all categories become active initially, /d/ indeed ends up as the only active category. This is also clear in Figure 8, which shows the network from Figure 6 after ten steps of activation spreading.

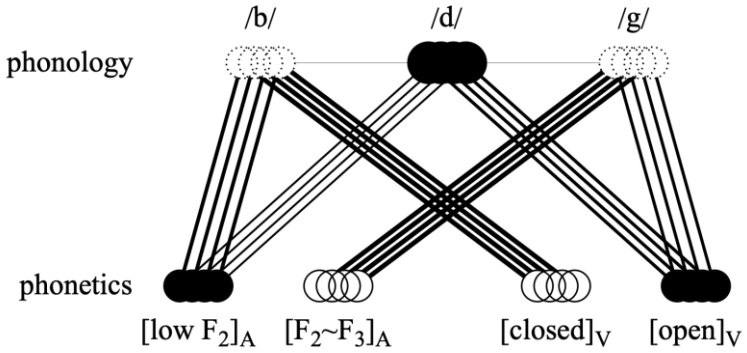


Figure 8: The network perceives auditory [ba] and visual [ga] as /da/.

In summary, when we feed the network auditory [ba] and visual [ga], it perceives /da/, replicating McGurk and MacDonald's observations. But what happens in the reverse effect, when we feed it auditory [ga] and visual [ba], the situation in which human listeners give combined responses such as /bagba/?

For auditory [ga] I deactivate the $[\text{low } F_2]_A$ nodes and activate the $[F_2{\sim}F_3]_A$ nodes; for visual [ba], I activate the $[\text{closed}]_V$ nodes and deactivate the $[\text{open}]_V$ nodes. From the shape of the mature network, we can already see what will happen next: since both groups of activated phonetic nodes are only connected to a single phonological category, both these categories become activated. The cumulative weights of the connections that lead to the categories are very nearly equal, namely 4.0000000; the difference between these sums is only $3.98 \cdot 10^{-8}$ (ca. 1 divided by 25 million). This minute difference plus the fact that the categories are activated in tandem, evidently leads to an equilibrium in which neither of the categories inhibits the other more than vice versa, as it happened in the regular McGurk effect. We therefore end up in an equilibrium with two winners: the network perceives both /b/ and /g/, as we can see in Figure 9.

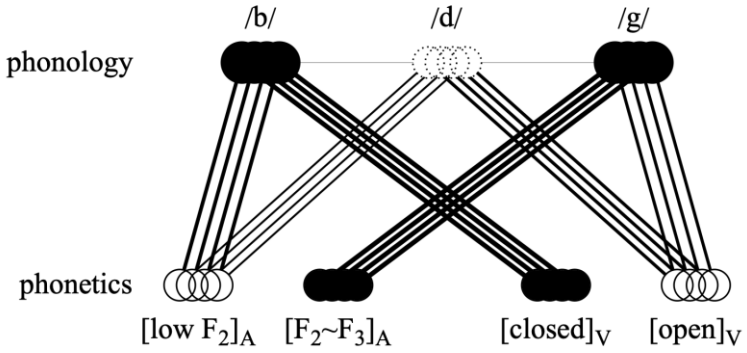


Figure 9: The network perceives auditory [ga] and visual [ba] simultaneously as /b/ and /g/.

This type of network is unable to represent temporal structure, so it does not possess any knowledge of phonotactic restrictions and therefore cannot express the language-specific knowledge that led English listeners to perceive /.bag.ba./ (with /g/ and /b/ in different syllables), as OT could.

4 The second neural network model of the McGurk effect (three levels, continuous phonetics)

Although the neural network from the previous section fared pretty well (the lack of structural constraints aside), it is still very OT-like, and we wanted to move away from OT to neural nets because we wanted to model how categories emerge from continuous input. In this section, I reuse a more extensive version of the neural network model, namely the one from Chládková (2014), my dissertation (Seinhorst 2021) and Boersma, Chládková & Benders (2022), containing three layers of representation: phonetics, phonology and meaning.

The phonetic layer consists of two groups of nodes, each representing a modality: one group represents the basilar membrane in the inner ear, on which our learner perceives the F_2 continuum, and the other group represents the visual modality, in which our learner perceives the lip aperture continuum. These continua have both been discretized into 48 nodes, and both are connected by cue connections to a dedicated group of nodes at the phonological layer.

These groups at the phonological layer both consist of six nodes; within a group, each node is connected to the other nodes by lateral inhibitory connections such as those discussed in section 3. In the previous model, these connections ensured competition in the perception process (cf. 3.3); in this layout, they also ensure that nodes become especially attuned to parts of the phonetic continua.

All nodes at the phonological level, in both layers, are connected to three groups of four nodes each at the meaning level. These groups refer to three objects in the external world, labeled “b”, “d” and “g” for the sake of convenience. Again, each group is connected to both other groups by inhibitory connections. As in the previous model, the maximum activity of unclamped nodes is 2.

In terms of Boersma’s BiPhon model, the phonological layer is probably a mix of the Surface Form and the Underlying Form; for the meaning layer I imagine a more central semantic layer where each group represents an object, so that the network is trained on very basic input pairs of objects and cues. However, with an artificially small language the distinction is hard to make. The underpinning idea is that the model has two levels in the external world, for which no representation needs to emerge (which in practice means that I can clamp the corresponding nodes and determine their activities) because they are tangible and/or measurable. Although the mapping from phonology to semantics, so from phonological words to a meaningful representation, is object recognition, in the

rest of this section I refer to it as perception, which is normally the mapping from a phonetic form to a surface form; in the context of this article, with three monophonemic objects, the term "recognition" seems ambitious.

The network in its initial state is shown in Figure 10:

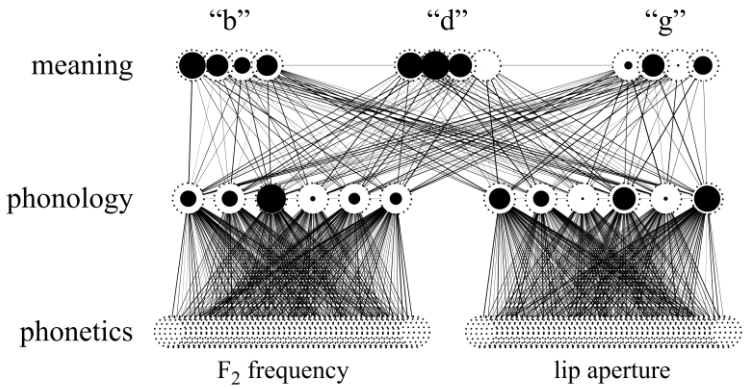


Figure 10: The three-level network in its initial state.

The auditory realizations of [b] and [d] must be fairly similar, with F_2s that are not far apart; [g], on the other hand, must be fairly distinct. As the network cannot represent temporal structure I cannot feed it formant transitions; instead, I choose to neglect F_3 and simplifyingly assume that F_2 is the only place cue in plosives. This way I can still maintain the auditory similarity of [b] and [d] versus [g] in a single continuum. I assume that the F_2 values of the auditory distributions are normally distributed, with the average [b] lying at 20% of the auditory continuum, the average [d] at 40%, and the average [g] at 80%; all distributions have a standard deviation that equals one-sixteenth of the continuum. The probability distribution of the auditory realizations, discretized into 48 nodes, then looks as follows:

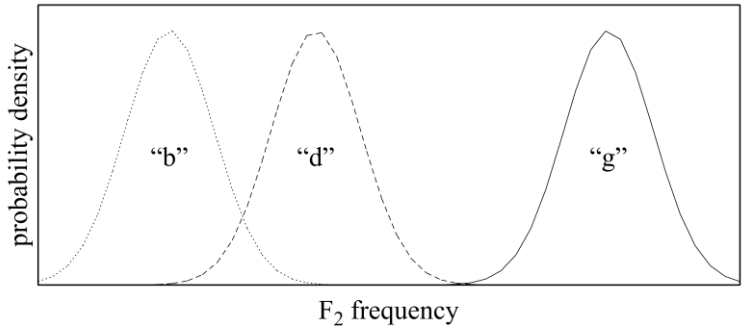


Figure 11: The probability densities of the F_2 frequencies of the realizations of the three categories.

In the visual modality, [b] is marked by a lack of visible lip aperture, while lip aperture can be seen in [d] and [g]; I assume their values to be normally distributed as well, with a standard deviation that equals one-tenth of the continuum. The mean lip aperture of [b] corresponds to 33.3% of the continuum, that of [d] and [g] to 66.7%. In reality, the continuum then probably corresponds to visible jaw height, and its leftmost tail to lips that are pressed together; for the sake of continuity, I will keep calling this dimension lip aperture.

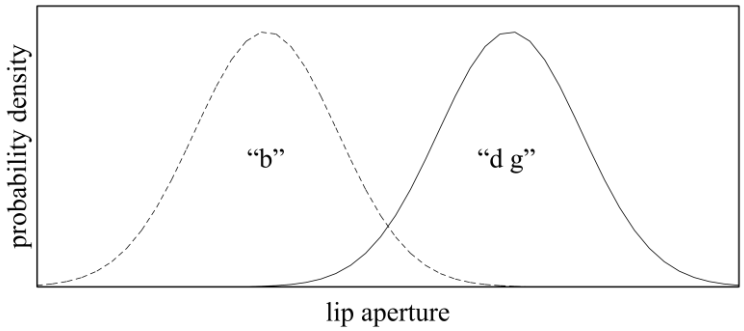


Figure 12: The probability densities of the lip apertures of the realizations of the three categories, discretized into 48 nodes: the curves of the "open" categories are identical.

4.1 The learning process

In this model, learning means to associate sensory cues with objects by forming intermediate representations in the phonology. As I did in my dissertation (Seinhorst 2021, section 2.5), I assume that the lexical representations unfold slowly; I implement this by having the activations at the meaning layer follow the logistic function drawn in Figure 13. Each simulation in this section consists of 10,000 learning steps. In the first learning step, the activation at the meaning layer is only marginally larger than 0; in the final step, it is only marginally smaller than 1.

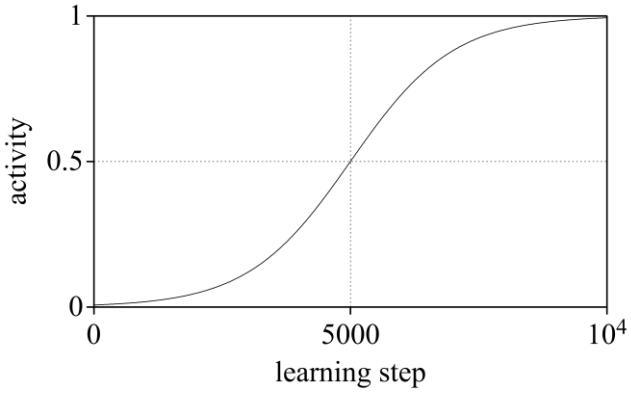


Figure 13: The activations at the meaning layer follow a logistic curve, assuming that the lexicon only unfolds slowly during the learning process.

Because the simulated child has practically no lexical representations initially, they learn from their sensory environment only. Because they cannot yet distinguish between categories, this environment is essentially the sum of the probability distributions from Figures 11 and 12. This pooled distribution in the auditory modality, discretized into 48 nodes, looks as follows:

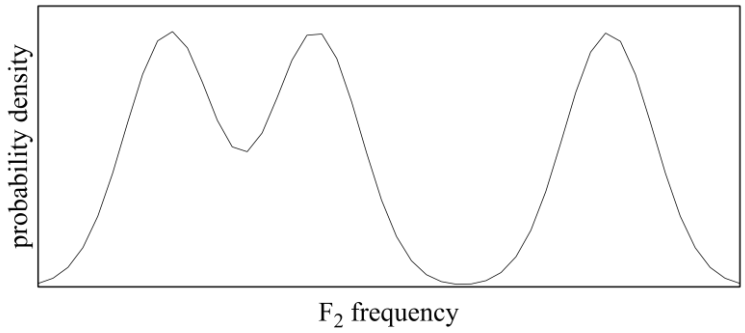


Figure 14: The pooled distribution of F2 frequencies in the learner's environment.

The pooled distribution of the lip apertures, discretized into 48 nodes, is shown in Figure 15: the right peak is twice as high as the left one, because the latter only contains [b] tokens, and the former contains both [d] and [g] tokens.

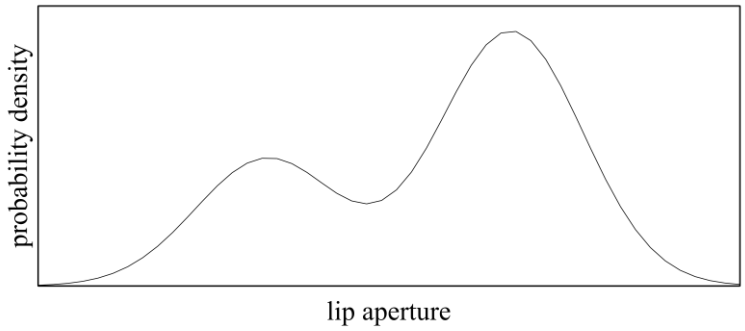


Figure 15: The pooled distribution of lip apertures in the learner's environment.

From these peaks and valleys, the learner will induce phonological categories.

Every learning step is composed of several processes. Firstly, a meaning category is selected at random, and for each phonetic continuum, a token is drawn from the probability distribution in Figures 11 and 12. Some transmission noise is added to both tokens, with a mean of 0 and a standard deviation of 5% of the continuum. Secondly, a Gaussian distribution is computed around each of the tokens, with a standard deviation of 1.5 node; in the auditory modality, this spreading reflects that adjacent hair cells on the basilar membrane become activated too, and in the visual modality it may signify jaw movement. The

computed activities are then applied to the nodes at the phonetic layer. Thirdly, the nodes of the selected category at the meaning layer are activated according to the logistic function from Figure 15 above, while the nodes of the other meaning categories are set to zero. Fourthly, the nodes at the phonetic and meaning layers are clamped, meaning that their activities remain fixed. Fifthly, activity is allowed to spread through the network and settle in 25 time steps; finally, the weights of the connections are updated according to the inoutstar rule. Figure 16 shows a possible first learning step. We see that the meaning nodes are clamped, but the activities are almost 0.

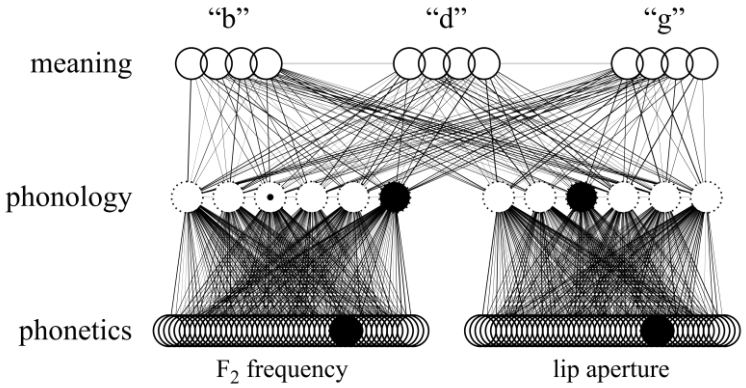


Figure 16: A learning step in the three-level neural network: the activities at the meaning and phonetic layers are fixed.

4.2 The halfway point: categories and production

After repeated exposure, patterns emerge in the phonological layer, where nodes become specialized in parts of both continua. Figure 17 shows the same network in its 5,000th learning step, acquiring a token of "d":

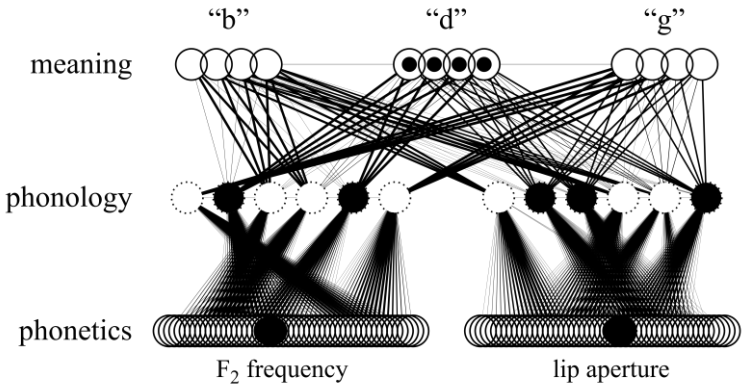


Figure 17: The same network after 5,000 learning steps. Categorical behavior is emerging in the phonology.

On the left phonological layer, it seems that nodes 1 and 6 have become connected to the higher F_2 frequency nodes exclusively, so the part of the continuum occupied by [g], and to the meaning nodes "g"; nodes 2 and 5 have become most

strongly connected to [d] and "d", while nodes 3 and 4 are most strongly connected to [b] and "b". Nevertheless, we see that nodes 2 and 5 are also weakly connected to [b] and "b", and nodes 3 and 4 are also somewhat connected to [d] and "d". At this point, then, this learner seems to represent place of articulation primarily, with a hint of a feature [low F_2] (and perhaps [high F_2], which coincides with and is therefore indistinguishable from [velar]).

On the right phonological layer, nodes 1 and 4 appear to be most strongly connected to the left half of the aperture continuum and to "b"; nodes 2, 3 and 6 appear to be connected to the right half of the continuum and to "d" and "g" but somewhat to "b" as well; and node 5 is connected to both sides but leans towards the left half of the continuum and towards "b". At this point, a binary contrast between [open] and [closed] seems to emerge.

Even though the activities at the meaning layer have not yet exceeded 0.5 at this point, the learner can already produce adult-like auditory and visual distributions, that is, distributions that match its input. When I ask it to produce "b", by first activating and clamping the "b" nodes at the meaning layer and spreading activity downwards to the phonology in 25 steps, then clamping the phonological nodes and spreading activity downwards to the phonetics in 25 steps, we see bell-shaped activity patterns. Their peaks lie around 20% of the F_2 frequency continuum and one third of the lip aperture continuum (see Figure 18); the realizations of "d" and "g" are also adult-like, with peaks at two thirds of the aperture continuum and 40 and 80 percent of the F_2 continuum, respectively (Figures 19 and 20). If we interpret these activities directly as probabilities, the resulting distribution strongly resembles the ambient language.

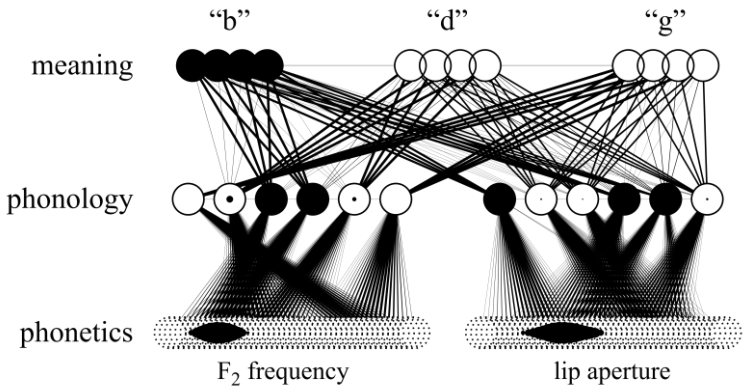


Figure 18: At the halfway point of acquisition, after 5,000 steps, the model produces adult-like activity patterns. In this figure, the network produces the "b" meaning category. The activities at the phonetic layer, which I interpret as output probabilities, are more or less bell-shaped, with their peak locations mirroring the peaks in their input language.

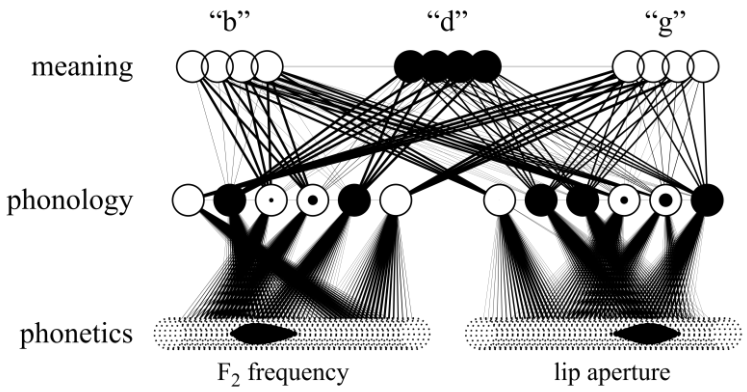


Figure 19: The network producing the "d" meaning category.

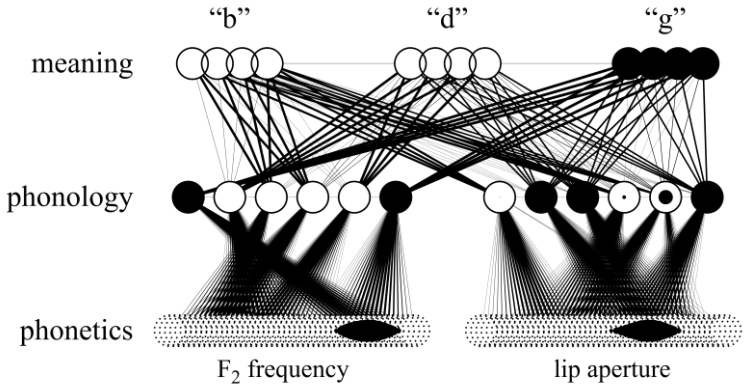


Figure 20: The network producing the "g" meaning category.

Although the BiPhon model is in essence parallel rather than serial, the process of asking the network to speak has to be sequential, or else we run into trouble: if we let activity flow freely, for instance in the production of "d", activating the feature [open] would also activate the "g" nodes on the meaning layer, which would then activate nodes on the left phonological layer, which would then activate its nodes on the F_2 continuum; and because "d" would activate nodes on the leftmost phonological layer that it shares with the /b/ category, these would activate "b", which would activate the feature [closed lips], et cetera. Finally, we would end up with a network in which most if not all unclamped nodes, so the nodes on the phonological and meaning layers, are activated. Letting the activity spread in two steps and keeping the phonological nodes clamped in the second step means that activity spreading remains contained. Perception will have to proceed in the same way, to avoid the same problem.

4.3 The mature network: categories

After 10,000 inputs, the connections in the network do not change anymore, so I consider the network stable and mature. The same network is shown in Figure 21.

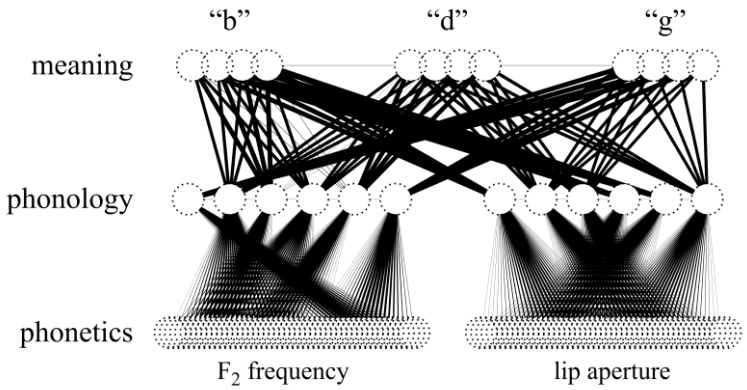


Figure 21: The mature network after 10,000 learning steps.

Comparing the phonological layer to its previous self after 5,000 steps, we see that on the leftmost side, nodes 1 and 6 are still connected to high F_2s and "g"; nodes 2 through 5, which were previously mostly (but not exclusively) connected to one part of the F_2 continuum and to a single meaning category, have now gone different ways. Nodes 2 and 4 are connected to "b" as well as "d" in apparently equal amounts, and to both peaks at the F_2 continuum, so they represent a [low F_2] feature; node 3 is mostly connected to "b" and its F_2 peak, but also a little bit to "d" and [d], while node 5 is mostly connected to "d" and [d] with weak connections to "b" and [b] as well, so these nodes seem to represent either place of articulation or the phonemes themselves.

On the right side of the phonological layer, nodes 1 and 4 are connected exclusively to "b" and [b], and node 5 has joined them unambiguously, shedding its connections to "d g" and [d g]; these nodes represent a [closed] feature. The other nodes encode a feature [open].

Repeated simulations show that the representations that this network has induced, mixing [low F_2] with either place of articulation or phoneme on the auditory side and a 3–3 contrast on the visual side, are by far the most common. 2–3 and 2–4 divisions in the right phonological layer occur as well.

4.4 The mature network: production and perception

When we ask the mature network to speak, we see that the output distributions of "b" and "d" have merged to a certain degree through the mediation of the feature [low F_2]; however, because both "b" and "d" have an additional more or less dedicated node, the overlap is not complete and both distributions retain their own peaks. Figures 22, 23 and 24 show the activity patterns for "b", "d" and "g" respectively.

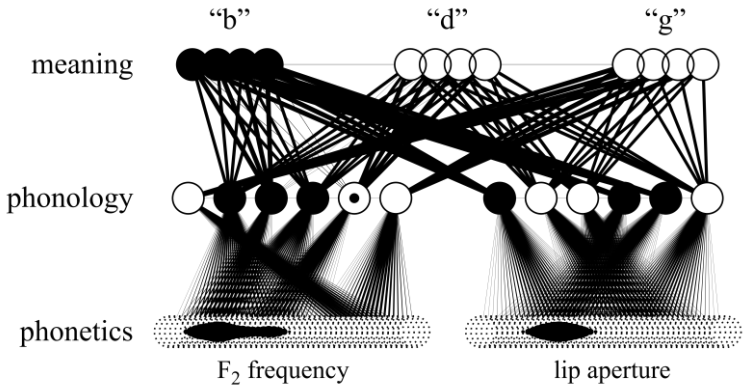


Figure 22: The mature network producing "b". As the feature [low F_2] has emerged, the output probabilities of [b] and [d] now show some overlap.

On the aperture continuum, we see that the distribution of the realizations of the [open] feature in Figures 23 and 24 is a bit wider than that of the [closed] feature: after all, there were twice as many input tokens of [open], so the network has encountered a larger number of peripheral realizations of [open] than of [closed] sounds.

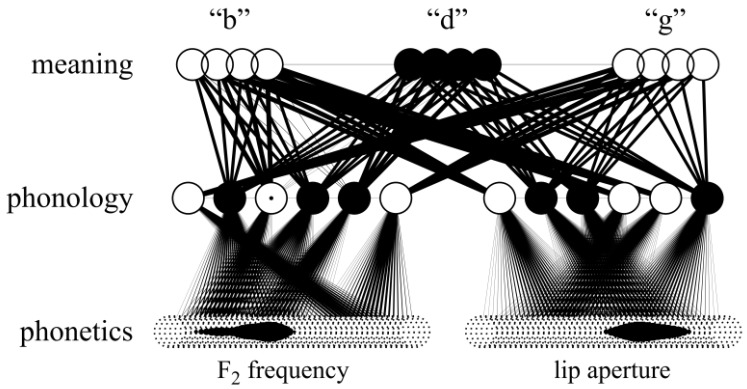


Figure 23: The mature network producing "d". As the feature [low F_2] has emerged, the output probabilities of [b] and [d] now show some overlap. The distribution of [open] is a bit wider than that of [closed] in Figure 22, because the network has encountered more extreme realizations of the open categories ([d g]) than of the closed one (only [b]).

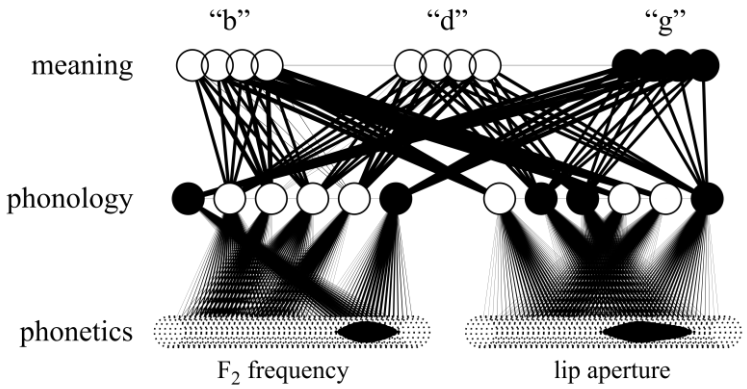


Figure 24: The mature network producing "g". The distribution of [open] is a bit wider than that of [closed] in Figure 22, probably because the network has encountered more extreme realizations of the open categories ([d g]) than of the closed one (only [b]).

Before I ask the model to perceive incongruent cues to mimic the McGurk effect, I want to be sure that the perception of congruent cues proceeds as expected. To do so, I compute a Gaussian distribution with the peak of the category as the mean (so 20% of the F_2 continuum and 33.3% of the aperture continuum for [b], 40% and 66.7% respectively for [d], and 80% and 66.7% respectively for [g]) and a standard deviation of 1.5 nodes to reflect the activity spreading on the basilar membrane; the maximum activity, at the mean, equals 1. On the phonetic layers, I apply the activities to the nodes and keep them clamped while keeping all other nodes unclamped, after which I let activity spread in 25 steps; after this, I clamp the nodes at the phonological layers and let activity spread again in 25 steps.

In the perception of [b], only "b" is activated, as expected:

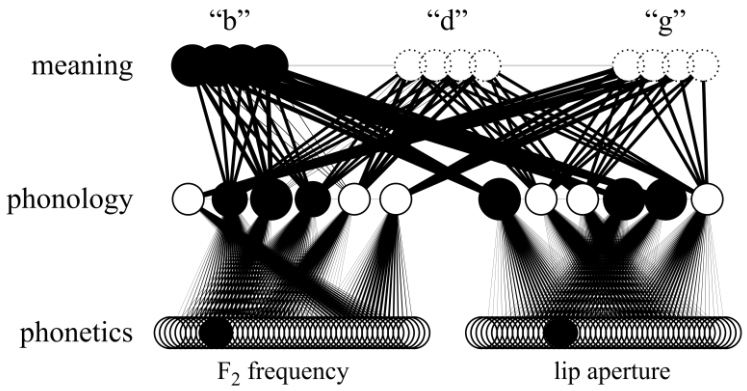


Figure 25: The network perceiving a token of [b] as "b", as expected.

However, in the perception of a token of [d], "g" becomes activated too:

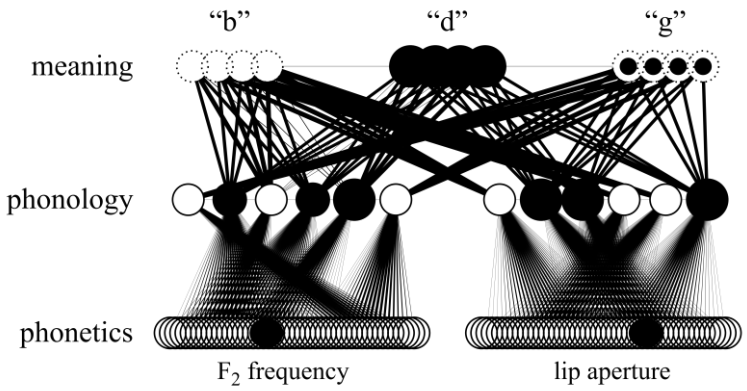


Figure 26: The network perceiving a token of [d]. Although the "d" category becomes most active, as desired, the "g" category becomes activated too.

Figure 27 shows that this activation is stable and not a result of the number of steps in which the activity settles in the network.

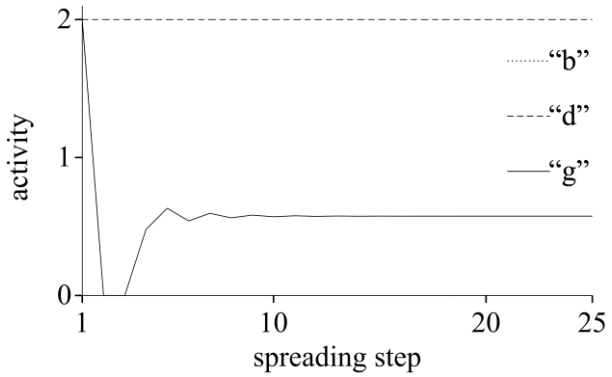


Figure 27: Activities in the meaning layer when the network perceives a token of [d]. Although the "d" category becomes most active, as desired, the "g" category becomes activated too.

This happens because the cumulative activity of the [open] nodes at the right phonological layer is higher than the cumulative activity of the [low F_2] and [coronal] (or |d|) nodes at the left phonological layer: this seer-listener prioritizes the visual modality (unlike the Japanese seer-listeners from Sekiyama and Tohkura 1991). Apparently, the lateral inhibition from "d" is not strong enough to nullify the activity in "g".

A possible solution to this undesired activation is to balance out the activities coming from the phonological layer by not allowing them to exceed 1. In this case, then, the cumulative activity coming from the left phonological layer will become 3, and so will the cumulative activity coming from the right phonological layer. This has the desired effect of only activating "d":

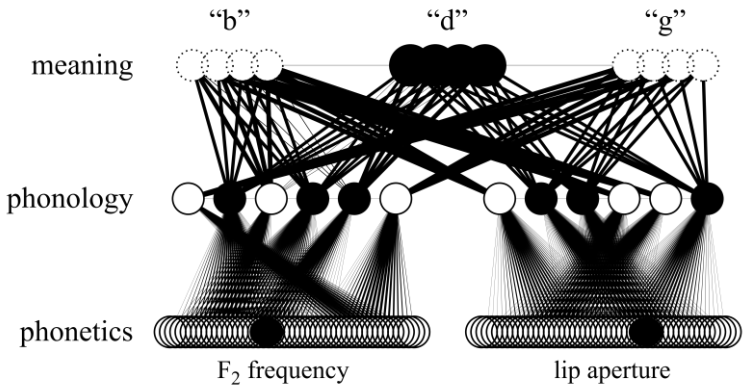


Figure 28: The network perceives a token of [d] as "d", if the activities coming from the phonological layers are equal.

This adjustment leaves the perception of [b] unaffected.

The perception of [g] proceeds as expected, yielding "g":

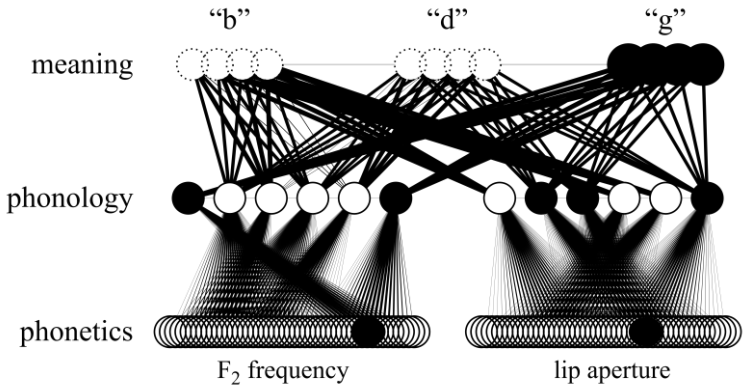


Figure 29: The network perceives a token of [g] as "g", as expected.

The [open] nodes will activate [d] too, but because the [dorsal] feature (or |g| phoneme) activates only "g" nodes, the lateral inhibitory connections from "g" cancel out the activity in "d". Although in this case, it is not necessary that the nodes at both phonological layers have the same cumulative activities (and in Figure 29, indeed they do not), it does not hurt either.

4.5 Does the McGurk effect happen?

Unlike the teleologically designed network in section 3, the model in this section is a general model of phonological and phonetic acquisition; it was not designed to replicate the McGurk effect. Rather, if the effect does turn up, it will be an unintended side effect of mechanisms and knowledge that are already in place in the perception of speech (as is true for the effect in human perceivers, of course). In this subsection I will put the network to the test.

In Figure 30, the network perceives auditory [b] together with visual [g]. The outcome of this process is “d”, as expected: the network can replicate the McGurk effect.

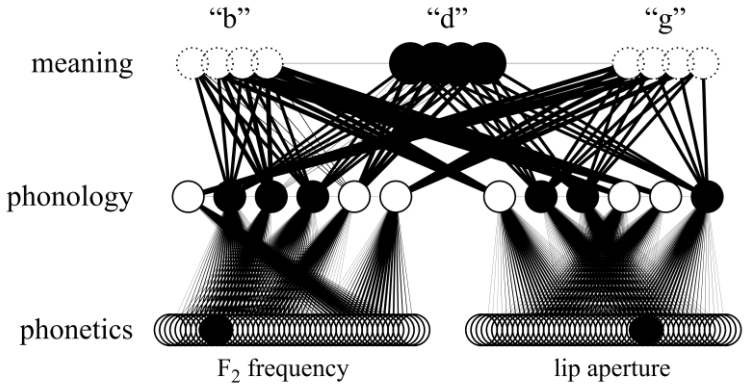


Figure 30: The network replicates the McGurk effect by perceiving a low F_2 token with open lips as “d”.

Figure 31 shows the activities in the meaning nodes as activity spreads through the network. Although the “b” and “g” nodes do become active initially (the corresponding lines coincide in the figure), the [low F_2] nodes that are activated by [b] and the [open] nodes that are activated by [g] also activate “d”, to a degree where its lateral inhibitory connections deactivate its competitors and it remains as the only candidate.

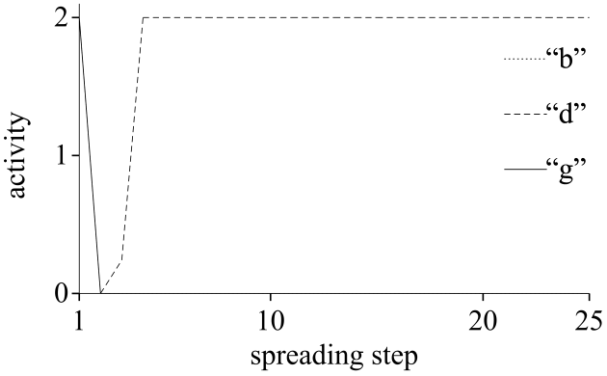


Figure 31: Activities at the meaning layer as a function of activity spreading step when perceiving auditory [b] with visual [g]. Initially both “b” and “g” become active, only to be overtaken by “d”, which remains active.

The next test is the reverse McGurk effect, where auditory [g] is paired with visual [b]. In this network, the perception process yields “b” as the outcome:

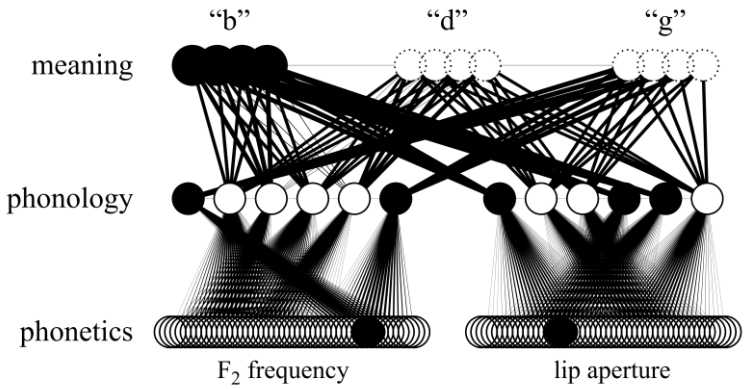


Figure 32: The network perceiving auditory [g] with visual [b] as "b", where human perceivers combine "b" and "g".

Based on McGurk and MacDonald's observations, we would expect the network to perceive both "b" and "g"; in fact, none of McGurk and MacDonald's participants perceived only "b". The problem here is again due to the fact that the cumulative activity at the right phonological layer ($\Sigma_a = 3$) exceeds that at the left layer ($\Sigma_a = 2$), simply because the number of active nodes is higher there: on the right layer, six nodes represent two peaks in the input distribution, while at the left layer, six nodes represent three peaks. As a result, the network again prioritizes vision over hearing.

If we balance the two senses by manually setting the activities of the active nodes at the auditory layer to 1.5, we achieve the desired result:

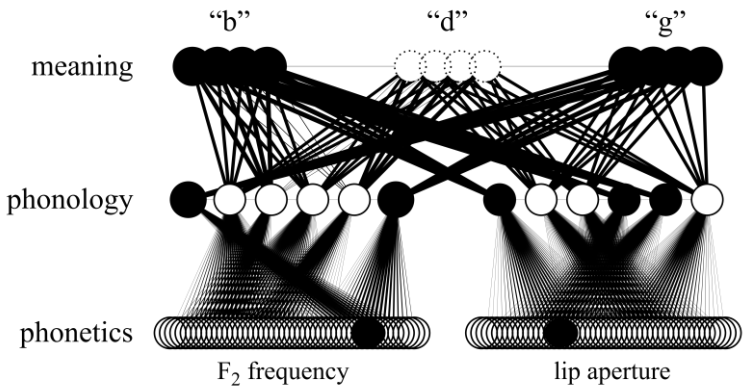


Figure 33: The network perceiving auditory [g] with visual [b] as simultaneous "b" and "g", as expected.

Because this network, like the one presented in section 3, is incapable of representing sequential information, I cannot comment on any phonotactic restrictions.

4.6 When does it work?

In the subsections above, I have established that the cumulative activities spreading upwards from the phonological layer must be identical for the results to make sense, or in other words, the auditory and visual modalities have to contribute equally, even in the perception of congruent cues. If this condition is not met, the resulting disbalance yields undesired effects, at least when we take McGurk and MacDonald's result as our objective, ranging from mixed outputs, as in Figure 26, to complete disregard of the modality with the weakest activation, as in Figure 32.

The simulations above relied on [b] being auditorily similar to [d], with their distributions being not only adjacent but close. How does the network fare if it cannot induce a [low F_2] feature?

To test this, I taught the network an ambient language in which the F_2 realizations of the meaning categories were evenly spaced on the continuum, with peaks at 20, 50 and 80 percent, as shown in Figure 34, so with the mean [b] and [d] being 30% apart instead of 20%. The distributions of the lip aperture continuum are unchanged.

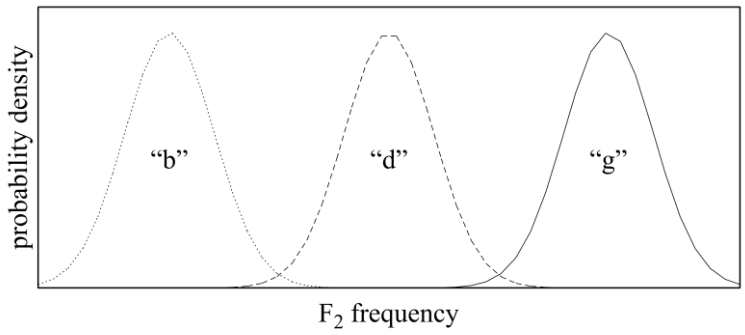


Figure 34: An input language in which the probability distributions of the auditory realizations of the three meaning categories are evenly spaced on the F_2 continuum.

The valleys between the peaks are deep enough for this network to induce three categories on the leftmost phonological layer. Figure 35 shows the mature network.

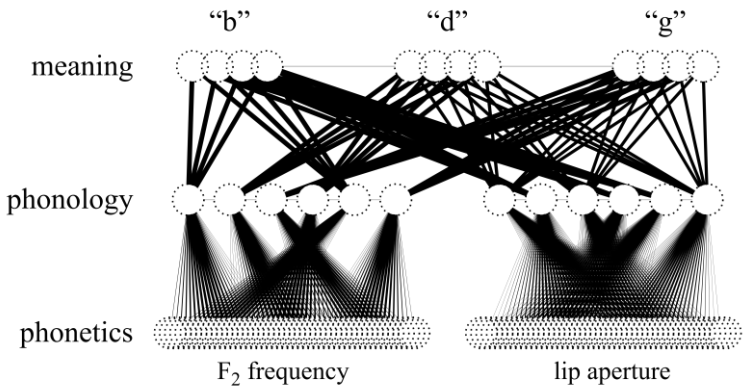


Figure 35: A network that has been exposed to the language from Figure 34. A ternary contrast has emerged on the left phonological layer, a binary contrast on the right.

On the left phonological layer, we see that nodes 1 and 5 have become specialized in the [b] tokens, nodes 2 and 4 in the [d] tokens, and nodes 3 and 6 in the [g] tokens; the network has encoded either a ternary place feature or three phonemes in the phonology, but there is no class with more than one member, as expected. On the right phonological layer, the familiar 3–3 division has emerged: nodes 2, 4 and 5 represent the [closed] feature, nodes 1, 3 and 6 represent the [open] feature.

Does this network still exhibit the McGurk effect? If, following the procedure described in 4.4 (below Figure 24), I apply activity at 20% of the F_2 continuum (the mean value for [b]) and at two thirds of the aperture continuum (the mean value for [g]), then spread activity upwards to the phonology, multiply the values of the activities at the left layer by 1.5 to make sure the cumulative activities of the two phonological layers are equal, and then spread activity upwards to the meaning layer, we see that "b" wins. (Without the multiplication, "d" wins, because the three [open] nodes outnumber the two [labial] nodes.)

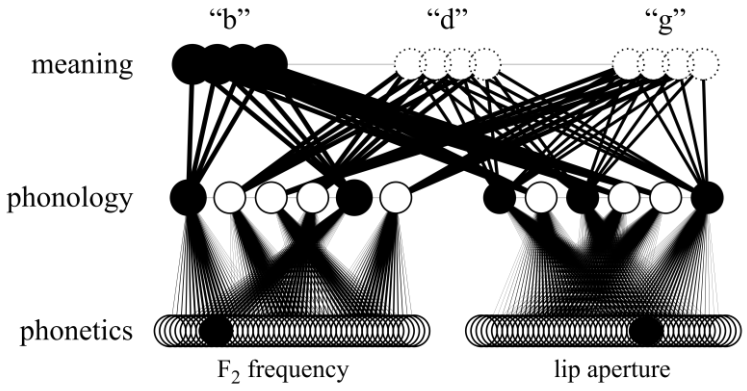


Figure 36: If [b] and [d] are further apart, the mean auditory F_2 of [b] together with the mean aperture of [g] elicit "b", not "d".

In the middle region of the F_2 continuum, both the auditory and the visual cues are congruent with "d", so we can find out at which point the network flips from "b" to "d" by raising F_2 in increments of one node. Figure 37 plots the activities at the "b" and "d" meaning nodes when the mean activated F_2 frequency lies between nodes 1 to 25. In all cases, the mean activated node on the lip aperture continuum is at two-thirds of the continuum, so the average [open] token. For the sake of comparability, Figure 38 plots the probability densities of [b] and [d] in the input language from F_2 nodes 1 to 25, so essentially it zooms in on Figure 34.

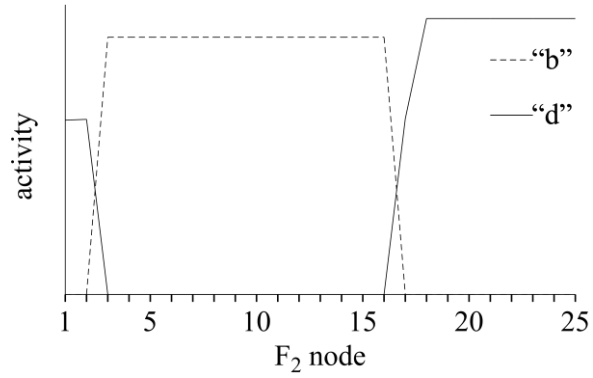


Figure 37: The activations of the "b" and "d" meaning nodes as a function of F_2.

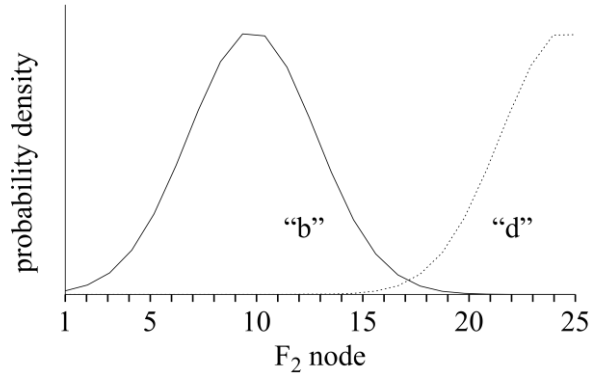


Figure 38: The probability densities of the F_2 realizations of "b" and "d" in the ambient language, as a function of F_2 node (showing nodes 1 to 25).

Figure 37 shows that in the lowest region of F_2, the perceiver surprisingly perceives "d", perhaps due to a lack of experience with such peripheral tokens; they then perceive "b" up to and including node 16, and starting from node 17, "d" is the most activated category. In Figure 38, we see that while this node is in the right tail of the distribution of [b] tokens, with an input probability still just above zero, it lies in the area where [b] and [d] overlap and confusion with [d] already exists. In a natural setting, the listener probably distinguishes such [b]s

from [d]s by a top-down influence from the lexicon, which plays no role in the meaningless syllables that elicit the McGurk effect.

It is unrealistic that the effect only happens in such a narrow band of F_2 frequencies; I surmise that the fragility of the effect in this simulation results from the oversimplification in the earlier simulations that F_2 is the only cue for place of articulation. We will probably need a more richly defined input language in which [b] and [d] are auditorily similar in one respect (but different from [g]), and another one in which they are auditorily different (also from [g]). If the properties are not isometric, the model would need to be extended with a third phonetic (i.e., second auditory) layer; however, the former property is likely the distance between F_2 and F_3 (small for [g], large for [b d]), and the latter property is F_2 locus (different for all three categories). In this case, both properties would activate nodes on the same continuum, as in Chládková's (2014) simulations of the emergence of vowel features, who demonstrated how features in both the height and backness dimensions emerge on the same basilar membrane. However, a crucial difference with vowel features is that in the perception of plosives, the distance between formants plays a central role (after all, one of the top-ranked constraints in Tableau 1 refers to it), but the network does not compute differences between formants; it creates categories around peaks, not around distances between peaks. In order to improve the model and its stability over multiple generations, these differences may need to be represented on a separate layer after all.

5 Conclusion and discussion

Taking Boersma's Optimality-Theoretic explanation of the McGurk effect as a starting point, I have presented two neural network models in which the effect arises: they replicate both the unifying percept /dada/ as a result of auditory [baba] and visual [gaga], and the combination percepts such as /bagba/ as a result of visual [baba] and auditory [gaga].

The first model consists of two layers, a phonetic and a phonological one, and the phonetic layer represents only discrete, given cues such as [low F_2] or [open lips]. This model is a translation of Boersma's formalization into a neural network and has thus been designed to replicate the effect, but it does provide an accessible visualization of how it emerges.

The second model has larger ecological validity, with three levels of representation: meaning, phonology and phonetics. The phonetic level consists of two layers representing continuous properties, namely (auditory) F_2 and (visual) lip aperture, and the learner induces phonological categories such as [open lips] based on the phonetic distributions (bottom-up) and the meaning categories (top-down). In this model, the McGurk effect is still present, as an unintended result

of phonological category learning. Both modalities need to contribute to the perception process equally, or else we get results that do not reflect the outcomes with English seer-listeners (but because there seem to be culture-specific elements to the effect, such a shifted balance would explain results from some non-English participants.)

Neither of the models incorporates sequential information, so the language-specific phonotactic restrictions that we have seen in human perceivers remain unexplored; and although the second model overcomes a lot of objections against Optimality Theory, it lacks the plosive formant transitions that cue place of articulation and might play a crucial role in the effect, or at least in the model's ability to induce place features.

While the effect does emerge in the second model, it requires a large degree of overlap between the input distributions of [b] and [d]; from these distributions, the network concurrently induces individual categories and a shared [low F_2] feature, the latter of which helps in mimicking the McGurk effect. As a result of the shared feature, the output distributions show a small but unrealistic overlap, and based on the simulations in my dissertation (Seinhorst 2021, section 3.5) we can expect such distributions to inevitably merge after a small number of generations. In a distribution with less overlap, in which the network does not induce a feature, the McGurk effect only occurs in peripheral [b] tokens in the right tail of the distribution, so in tokens that are prone to confusion with [d] anyway. The solution probably lies in a more refined input on the auditory level, ideally causing the network to represent both F_2 locus (different for [b], [d], and [g]) and F_2–F_3 difference (larger for [b d] than for [g]) in the phonology, as a ternary and a binary contrast, respectively. Such a representation will probably keep the effect intact and remain stable over the generations.

6 References

Başkent, D., & Bazo, D. (2011). Audiovisual asynchrony detection and speech intelligibility in noise with moderate to severe sensorineural hearing impairment. *Ear and Hearing, 32*(5): 582–595. https://doi.org/10.1097/aud.0b013e31820fca23

Boersma, P. (2006/2012). A constraint-based explanation of the McGurk effect. In B. Botma & R. Noske (Eds.), *Phonological explorations: Empirical, theoretical and diachronic issues* (pp. 299–312). De Gruyter. Earlier version: https://www.fon.hum.uva.nl/paul/papers/McGurk.pdf

Boersma, P. (2011). A programme for bidirectional phonology and phonetics and their acquisition and evolution. In A. Benz & J. Mattausch (Eds.), *Bidirectional Optimality Theory* (pp. 33–72). John Benjamins. https://doi.org/10.1075/la.180

Boersma, P., Benders, T., & Seinhorst, K.T. (2020). Neural networks for phonology and phonetics. *Journal of Language Modelling 8*(1): 103–177. https://doi.org/10.15398/jlm.v8i1.224

Boersma, P., Chládková, K., & Benders, T. (2022). Phonological features emerge substance-freely from the phonetics and the morphology. *Canadian Journal of Linguistics/Revue canadienne de linguistique 67*(4): 611–669. https://doi.org/10.1017/cnj.2022.39

Boersma, P., & Hamann, S. (2008). The evolution of auditory dispersion in bidirectional constraint grammars. *Phonology 25*(2): 217–270. https://doi.org/10.1017/S0952675708001474

Boersma, P., & Hamann, S. (2009). Loanword adaptation as first-language phonological perception. In A. Calabrese & W. L. Wetzels (Eds.), *Loanword phonology* (pp. 11–58). John Benjamins. https://doi.org/10.1075/cilt.307.01cal

Burnham, D., & Dodd, B. (2004). Auditory–visual speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology 45*(4): 204–220. https://doi.org/10.1002/dev.20032

Chládková, K. (2014). *Finding phonological features in perception* [Doctoral dissertation, University of Amsterdam]. https://www.fon.hum.uva.nl/katerina/documents/chladkova_PhDthesis.pdf

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological Cybernetics 23*: 121–134. https://doi.org/10.1007/978-94-009-7758-7_12

Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science 11*(1): 23–63. https://doi.org/10.1016/S0364-0213(87)80025-3

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature 264*: 746–748. https://doi.org/10.1038/264746a0

Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience 25*(25): 5884–5893. https://doi.org/10.1523/JNEUROSCI.0896-05.2005

Prince, A., & P. Smolensky (1993/2004). *Optimality Theory: Constraint interaction in Generative Grammar*. Technical Report TR-2, Rutgers University Center for Cognitive Science. https://roa.rutgers.edu/files/537-0802/537-0802-PRINCE-0-0.PDF

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science 9*(1): 75–112. https://doi.org/10.1016/S0364-0213(85)80010-0

Seinhorst, K. T. (2009). *De* [y] *in Russische leenwoordadaptatie* [BA thesis, University of Amsterdam]. https://www.fon.hum.uva.nl/archive/2009/2009-BA-KlaasSeinhorst.pdf

Seinhorst, K. T. (2012). *The evolution of auditory dispersion in symmetric neural nets* [MA thesis, University of Amsterdam]. https://www.fon.hum.uva.nl/archive/2012/2012-MA-KlaasSeinhorst.pdf

Seinhorst, K. T. (2021). *The learnability of phonological patterns: simulations, experiments, typology* [Doctoral dissertation, University of Amsterdam]. LOT. https://dx.medra.org/10.48273/LOT0584

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory–visual speech perception. *Developmental Science 11*(2): 306–320. https://doi.org/10.1111/j.1467-7687.2008.00677.x

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America 90*(4): 1797–1805. https://doi.org/10.1121/1.401660

Traunmüller, H., & Öhrström, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics 35*(2): 244–258. https://doi.org/10.1016/j.wocn.2006.03.002

Valkenier, B., Duyne, J. Y., Andringa, T. C., & Baskent, D. (2012). Audiovisual perception of congruent and incongruent Dutch front vowels. *Journal of Speech, Language, and Hearing Research 55*(6): 1788–1801. https://doi.org/10.1044/1092-4388(2012/11-0227)

Klaas T. Seinhorst

University of Amsterdam, Department of Psychology

seinhorst@uva.nl