Integrating prediction errors at two time scales permits rapid recalibration of speech sound categories

Itsaso Olasagasti[1*], Anne-Lise Giraud[1]

[1]Department of Basic Neuroscience, University of Geneva, Geneva, Switzerland
[*]Corresponding author
E-mail: miren.olasagasti@unige.ch

Abstract

Speech perception can be derived from internal models of how speech sounds are systematically associated with sensory features. When features associated with a speech sound change, the listener should recalibrate its internal model by appropriately weighing new versus old evidence in a volatility dependent manner. Models of speech recalibration have classically ignored volatility. Models that explicitly consider volatility have been designed to describe the behavior of human participants in tasks where sensory cues are associated with arbitrary experimenter defined categories or rewards. In that setting, a model that maintains a single representation of the category but continuously adapts the learning rate works well. We argue that recalibration of existing "natural" categories is better described by a model that represents sound categories at different time scales. We illustrate our proposal by modeling the rapid recalibration of speech categories reported by Lüttke et al. (2016).

Introduction

The way the brain processes sensory information to represent the perceived world is flexible and varies in response to changes in the stimulus landscape. Neurocognitive adaptation to varying stimuli can be driven by an external feedback signal, but might also take place with simple passive exposure to a changing environment. In the domain of speech perception, sound categories are susceptible to such stimulus-driven recalibration. Typically, ambiguous speech stimuli that are disambiguated by context (McQueen, Cutler, and Norris 2006) or by a concurrent stimulus presented e.g. in the visual modality (Bertelson, Vroomen, and De Gelder 2003; Vroomen et al. 2007), are subsequently perceived as a function of the disambiguated percept even when presented alone.

Studies using a two alternative forced choice show that changes in acoustic speech categories can be induced by input from the visual modality (e.g. (Bertelson, Vroomen, and De Gelder 2003)). This effect is reciprocal; acoustic information can disambiguate lipreading (Baart and Vroomen 2010) and result in measurable aftereffects. While such effects can be observed after repeated exposure to the adapting stimuli, recalibration can occur quickly (Vroomen et al. 2007). This fast and dynamic process has been modelled as incremental Bayesian updating (Kleinschmidt and Jaeger 2015) where internal perceptual categories are modified to better match the stimulus statistics. In their Bayesian approach, when listeners are confronted with modified versions of known speech categories, they modify the perceived category's representation so as to make it more consistent with the features presented in the stimulus. The resulting recalibration weighs all evidence equally, independent of their recency. This model successfully describes perceptual changes observed when listeners are confronted with the repeated presentation of a single modified version of a speech sound in a blocked experimental design. However, it cannot appropriately deal with intrinsically changing environments, in which the relevance of sensory cues decreases with time, with a faster decrease the more volatile the environment.

Inference in volatile environments has been studied mostly in relation to decision making tasks in which participants-who receive explicit feedback- need to keep track of the varying statistics of arbitrary cue reward associations (Behrens et al. 2007) or arbitrarily defined categories (Summerfield et al. 2011). This work has shown that humans are able to adjust their learning rate

to the volatility in the stimulus set, showing faster learning rates in more volatile environments and vice versa -faster learning rates translate into a stronger devaluation of past evidence-. This has lead to normative models in which task relevant features are represented at a single although variable time scale, with the focus of the models being on the online estimation of volatility (Behrens et al. 2007; C. Mathys et al. 2011).

Similar variable learning rates might apply in speech. However, we would like to propose that when recalibrating existing "natural" categories, a normative model should account for the fact that these categories can change at different time scales. Transient changes in how a speech category sounds, such as those coming from a new speaker, should not interfere with long-term or more slowly varying representations of the same category that perhaps encode commonalities across speakers. We therefore propose that speech sound categories should be represented by more than a single varying timescale.

Although speech category recalibration has not been systematically studied in variable environments (Lüttke et al. 2016) found recalibration in an experiment that included McGurk stimuli in the absence of an explicit adapting condition, and even for acoustic stimuli that were not ambiguous. In their study, which involved six different vowel/consonant/vowel stimuli presented in random order, the fact that the acoustic stimulus /aba/ was mostly perceived as an illusory /ada/ when presented with the video of a speaker producing /aga/ (McGurk effect), was powerful enough to yield observable adaptive effects across consecutive trials; the probability of an acoustic /aba/ to be categorized as /ada/ rather than /aba/, was higher when the trial was preceded by an audio-visual McGurk fusion. Other work has shown that similar recalibration effects did not affect other phonetic contrasts/categories. After participants had recalibrated acoustic sounds in the /aba/-/ada/ continuum (in which formant transitions are the main acoustic cue), the recalibration did not generalize to /ibi/-/idi/ (main acoustic cue burst and frication), to /ama/-/ana/ (also cued by formant transitions) or to /ubu/-/udu/ (same phoneme contrast also relying on formant transitions (Reinisch et al. 2014). Similarly, in a word recognition task participants were able to keep different F0/VOT correlation statistics for different places of articulation (Idemaru and Holt 2014). We hence hypothesized that changes occur at the level of internal speech categories (the sublexical /ada/ category) rather than that of sensory features.

To test our hypothesis we simulated Lüttke et al.'s (Lüttke et al. 2016) experiment, using a model of audiovisual integration based on hierarchical Bayesian inference; we explicitly included a model for the McGurk effect based on our previous work (Olasagasti, Bouton, and Giraud 2015) and supplemented it with an adaptation mechanism that uses residual prediction error to update the internal representation – in the form of a hierarchical generative model- associated with the perceived category. The model divides the process in two steps; the first is an on-line (moment-by-moment) prediction error minimization during the perceptual inference process, at the end of which there might remain "residual" prediction errors. This is typically the case in fused McGurk stimuli, since the best explanation for the multisensory input (/ada/) is neither the audio /aba/ nor the visual /aga/, which leaves "residual" prediction errors in both acoustic and visual dimensions.

We were able to reproduce the results of Lüttke et al. (Lüttke et al. 2016) described above only when we considered that adaptation occurred at two different time scales resulting in 1) a transient effect leading to recalibration towards the most recently presented stimulus features, which decayed towards 2) a longer-lasting representation corresponding to a mapping between category and stimulus features determined within a longer time span. Interestingly, a cascade model of synaptic modification seems to share some of these features (Iigaya 2016).

Materials and Methods

The goal of inference is to establish which is the speech token that gave rise to the incoming sensory input. For our present purposes we restrict ourselves to three tokens /aba/, /ada/ and

/aga/ (as in (Lüttke et al. 2016)). Although several acoustic and visual features can distinguish between them, we choose to model the 2nd formant transition, which is minimal for /aba/ but increases for /ada/ and /aga/, and the degree of lip closure, which is maximal for /aba/ and less prominent for /ada/ and /aga/ (lip closure /aba/>/ada/>/aga/). This choice is based on the fact that what distinguishes between the three speech sounds is the place of articulation. Acoustically the 2nd formant transition is an important cue for place of articulation, particularly within the 'a' vowel context (LIBERMAN et al. 1957); visually it is the degree of lip aperture at the time of the vocal cavity occlusion depending on its location (complete lip closure for the bilabial (/aba/), and decreasing lip closure for the alveolar (/ada/) and velar /aga/) (Campbell 2008; Varnet et al. 2013).

The generative model has three levels; the higher level encodes the speech token, the speech token in turn determines the expected values for the audiovisual cues, as represented in Fig 1A. The model includes the three possible tokens, each determining the expected distribution of its associated audiovisual features. We also introduce uncorrelated noise to account for sensory variability. The parameters, location and spread of features associated with each token, as well as the level of sensory noise, define an individual listener's internal model.

Although not strictly necessary for our present purpose, the model includes the temporal profiles of the two sensory features (Fig 1B). The latter are fixed and only their amplitude is modelled as a random variable that changes only across trials. The inclusion of the temporal dimension allows us to gain intuition into the separation of the perceptual and recalibration processes.

We use 'k' as the speech token index k = {/aba/, /ada/, /aga/}; 'f' as feature index f = {V,A}, where 'V' stands for the visual feature (lip aperture) and 'A' for the acoustic feature (2nd formant transition). The idea is that the lip aperture and 2nd formant have stereotypical temporal modulation profiles that vary in amplitude according to the speech token.

We use '$M_{f,n}$' to denote the temporal profile of the modulation for feature 'f' at time point 'n' ($M_{V,1}$ … $M_{V,N}$) stands for the temporal profile of lip modulation and ($M_{A,1}$ … $M_{A,N}$) for the 2nd formant modulation profile. The time dependent modulation for each feature, lip '$y_{V,n}$:($y_{V,1}$ … $y_{V,N}$)' and 2nd formant '$y_{A,n}$:($y_{A,1}$ … $y_{A,N}$)' is thus modelled as the product of a constant modulation amplitude 'C_V' or 'C_A' and a standard temporal profile (($M_{V,1}$ … $M_{V,N}$) or ($M_{A,1}$ … $M_{A,N}$)). In this way, we are explicitly separating 'what' and 'when'. Thus '$y_{f,n}$' represents feature 'f' at time point 'n'. The full generative model linking a token 'k' to audiovisual sensory features is described by the equations below.

There are two sources of variability, moment-to-moment uncorrelated sensory noise (σ_V, σ_A) and the variability of modulation amplitudes associated with the same token in different realisations ($\sigma_{k,V}$, $\sigma_{k,A}$) k = {/aba/, /ada/, /aga/}.

The hierarchical generative model is defined by the following relations:

$$p(y_{f,n}|C_f) \propto \exp\left(-\frac{(y_{f,n}-C_f M_{f,n})^2}{2\sigma_f^2}\right) \tag{1}$$

$$\equiv \exp\left(-\frac{(s_{f,n}-C_f)^2}{2\sigma_{f,n}^2}\right), \quad s_{f,n} \equiv y_{f,n}/M_{f,n}, \quad \sigma_{f,n} \equiv \sigma_f/M_{f,n} \tag{2}$$

$$p(s_{f,n}|C_f) \propto \exp\left(-\frac{(s_{f,n}-C_f)^2}{2\sigma_{f,n}^2}\right), \quad f = A, V \tag{3}$$

$$p(C_f|k) \propto \exp\left(-\frac{(C_f-\theta_{k,f})^2}{2\sigma_{k,f}^2}\right), \quad f = A, V, \quad k = \text{speech token} \tag{4}$$

In the following we will work with '$s_{f,n}$' instead of '$y_{f,n}$', that is, with the information about the – constant- amplitude instead of the time varying feature profile. This makes the associated variance, $(\sigma_{f,n})^2$, dependent on the standard temporal profile of the relevant feature.

While the above defines the generative (top-down model) $p(s_{f,n}|C_f,k)$, our interest lies in its inversion $p(k|s_{f,1}, \ldots, s_{f,N})$, where $s_{f,1}, \ldots, s_{f,N}$ represents the sensory input in a single trial.

From the above equations one obtains:

$$p(k|s_{V,1}, s_{A,1} \cdots s_{V,n}, s_{A,n}) \propto \exp\left(-\frac{(\bar{s}_{V,n} - \theta_{k,V})^2}{2(\bar{\sigma}^2_{V,n} + \sigma^2_{k,V})} - \frac{(\bar{s}_{A,n} - \theta_{k,A})^2}{2(\bar{\sigma}^2_{A,n} + \sigma^2_{k,A})}\right) p(k) \tag{5}$$

$$\frac{\bar{s}_{f,n}}{\bar{\sigma}^2_{f,n}} \equiv \sum_{m=1}^{n} \frac{s_{f,m}}{\sigma^2_{f,m}}, \quad \frac{1}{\bar{\sigma}^2_{f,n}} \equiv \sum_{m=1}^{n} \frac{1}{\sigma^2_{f,m}}, \quad f = V, A \tag{6}$$

which results from marginalizing over 'C_V' and 'C_A'-the intermediate stages that encode the visual and acoustic features, explicitly:

$$\begin{aligned}
p(k|s_{V,1}, s_{A,1} \cdots s_{V,n}, s_{A,n}) &= \int dC_V dC_A \quad p(k, C_V, C_A|s_{V,1}, s_{A,1} \cdots s_{V,n}, s_{A,n}) \\
&\propto \int dC_V dC_A \quad p(s_{V,1} \cdots s_{V,n}|C_V) p(C_V|k) p(s_{A,1} \cdots s_{A,n}|C_A) p(C_A|k) p(k) \\
&\propto \exp\left(-\frac{(\bar{s}_{V,n} - \theta_{k,V})^2}{2(\bar{\sigma}^2_{V,n} + \sigma^2_{k,V})} - \frac{(\bar{s}_{A,n} - \theta_{k,A})^2}{2(\bar{\sigma}^2_{A,n} + \sigma^2_{k,A})}\right) p(k)
\end{aligned}$$

$$\int dC_V \quad \frac{1}{\sigma_{k,V}} \exp\left(-\frac{(C_V - C_{V,n,k})^2}{2\sigma^2_{V,n,k}}\right) \int dC_A \quad \frac{1}{\sigma_{k,A}} \exp\left(-\frac{(C_A - C_{A,n,k})^2}{2\sigma^2_{A,n,k}}\right)$$

$$\frac{C_{f,n,k}}{\sigma^2_{f,n,k}} \equiv \frac{\bar{s}_{f,n}}{\bar{\sigma}^2_{f,n}} + \frac{\theta_{k,f}}{\sigma^2_{k,f}}, \quad \frac{1}{\sigma^2_{f,n,k}} \equiv \frac{1}{\bar{\sigma}^2_{f,n}} + \frac{1}{\sigma^2_{k,f}}, \quad f = V, A$$

We assume that initial variance and prior probabilities are equal across categories (p(k)=1/3).

Alternatively, marginalization over 'k' gives the probabilities over the hidden variables 'C_V' and 'C_A', which we associate with encoding of stimulus features (lip aperture, 2nd formant transition) in visual and auditory cortex respectively.

$$p(C_V, C_A|s_{V,1}, s_{A,1} \cdots s_{V,n}, s_{A,n}) =$$

$$\sum_k p(k|s_{V,1}, s_{A,1} \cdots s_{V,n}, s_{A,n}) \prod_{f=V,A} \frac{1}{\sqrt{2\pi}\sigma_{f,n,k}} \exp\left(-\frac{(C_f - C_{f,n,k})^2}{2\sigma^2_{f,n,k}}\right)$$

$$\frac{C_{f,n,k}}{\sigma^2_{f,n,k}} \equiv \frac{\bar{s}_{f,n}}{\bar{\sigma}^2_{f,n}} + \frac{\theta_{k,f}}{\sigma^2_{k,f}}, \quad \frac{1}{\sigma^2_{f,n,k}} \equiv \frac{1}{\bar{\sigma}^2_{f,n}} + \frac{1}{\sigma^2_{k,f}}, \quad f = V, A$$

This shows explicitly how internal estimates of sensory features (lip aperture 'C_V' and 2nd formant 'C_A') are driven by bottom up sensory evidence (s_V, s_A) and top-down empirical priors, with each category 'k' contributing according to its internal expectations ($\theta_{k,V}$ $\theta_{k,A}$). When there is strong evidence for a given category 'k', the sum can be approximated by a single Gaussian centred at a

compromise between $\theta_{k,V}$ and s_V for the visual feature and between $\theta_{k,A}$ and s_A for the acoustic feature.

The centre and variance of each Gaussian can also be written in terms of moment-to-moment incremental updates

$$\frac{C_{f,n,k}}{\sigma^2_{f,n,k}} = \frac{s_{f,n}}{\sigma^2_{f,n}} + \frac{C_{f,n-1,k}}{\sigma^2_{f,n-1,k}}, \quad \frac{1}{\sigma^2_{f,n,k}} = \frac{1}{\sigma^2_{f,n}} + \frac{1}{\sigma^2_{f,n-1,k}}, \quad f = A, V$$

where *n* indexes the time point within the trial and *k* denotes the speech token. Or equivalently for the $C_{f,n,k}$:

$$C_{f,n,k} = C_{f,n-1,k} + \frac{\sigma^2_{f,n-1,k}}{\sigma^2_{f,n} + \sigma^2_{f,n-1,k}}(s_{f,n} - C_{f,n-1,k}), \quad f = A, V$$

which explicitly shows how the internal representation of the modulation magnitudes is updated by moment-to-moment prediction errors.

In principle the model could also be made to perform causal inference (Magnotti and Beauchamp 2017), that is, decide whether the two sensory streams belong to the same source and therefore should be integrated, or whether the two sensory streams do not belong to the same source, in which case the participant should ignore the visual stream. Since Lüttke et al. explicitly selected the participants that consistently reported /ada/ for the McGurk stimulus, these subjects were fusing the two streams. We hence assume that integration is happening at every audio-visual trial.

The model's percept corresponds to the category that maximizes the posterior distribution $p(k|s_{f,1}, \ldots, s_{f,N})$.

Recalibration model. The previous section presented how the model does inference in a single trial. We now turn to how the model updates the parameters that encode the internal representation of the three speech categories. This happens after every trial, thus simulating an internal model that continuously tries to minimize the difference between its predictions and the actual observations; we assume that in this process, in which the model tries to make itself more consistent with the input just received, it will only update the category corresponding to its choice. We will present three updating rules. The normative incremental Bayesian update model used by (Kleinschmidt and Jaeger 2015), and the empirically motivated constant delta rule and hierarchical delta rule with intrinsic decay.

Bayesian updating. The internal representation of the speech categories is determined by six location parameters ($\theta_{k,V}$ $\theta_{k,A}$) and six width parameters ($\sigma_{k,V}$, $\sigma_{k,A}$). We follow Gelman et al. (Gelman et al. 2003) and define the following prior distributions for the internal model parameters ($\theta_{k,V}$, $\sigma_{k,V}$; $\theta_{k,A}$, $\sigma_{k,A}$). For each of the six (θ, σ) pairs (2 sensory features × 3 categories) the prior is written as:

$$p(\theta_{k,f}, \sigma_{k,f}) = p(\theta_{k,f}|\sigma_{k,f})p(\sigma_{k,f})$$
$$\propto \sigma^{-1}_{k,f}\left(\sigma^2_{k,f}\right)^{-(\nu_{k,f,0}/2+1)} \exp\left(-\frac{[\kappa_{k,f,0}(\theta_{k,f} - \mu_{k,f,0})^2 + \nu_{k,f,0}\sigma^2_{k,f,0}]}{2\sigma^2_{k,f}}\right)$$

As above, f refers to the sensory feature, either V or A, and k to the speech category, either /aba/, /ada/ or /aga/.

After a new trial with sensory input (s_V, s_A) the updated prior has the following parameters:

$$
\begin{aligned}
\kappa_{k,f,0} &\leftarrow \kappa_{k,f,0} + 1 \\
\mu_{k,f,0} &\leftarrow \mu_{k,f,0} + \frac{1}{\kappa_{k,f,0} + 1}(\bar{s}_f - \mu_{k,f,0}) \\
\nu_{k,f,0} &\leftarrow \nu_{k,f,0} + 1 \\
\nu_{k,f,0}\sigma^2_{k,f,0} &\leftarrow \nu_{k,f,0}\sigma^2_{k,f,0} + \frac{\kappa_{k,f,0}}{\kappa_{k,f,0} + 1}(\bar{s}_f - \mu_{k,f,0})^2
\end{aligned}
$$

After each trial the inference process described in the previous section determines the percept from the posterior probability $p(k|s_V\ s_A)$. Only the feature parameters of the representation corresponding to the percept are subsequently updated.

We use the values that maximize the posterior over the parameters given the input and the current estimated category 'k' to determine the point estimates that will define the updated model parameters for the next trial (Eq 4). The updates for the location and spread parameters then take the form:

$$
\begin{aligned}
\Delta\theta_{k,f} &= \frac{1}{\kappa_{k,f,0} + n(k)}(\bar{s}_f - \theta_{k,f}) \\
\Delta\sigma^2_{k,f} &= \frac{1}{\nu_{k,f,0} + n(k) + 1}\left[\frac{\kappa_{k,f,0} + n(k) - 1}{\kappa_{k,f,0} + n(k)}(\bar{s}_f - \theta_{k,f})^2 - \sigma^2_{k,f}\right]
\end{aligned}
\tag{7}
$$

Where 'k' is the perceived category, n(k) the number of times the category has been perceived, f =V,A designates the sensory feature and $\nu_{k,f,0}$ and $\kappa_{k,f,0}$ are parameters from the prior distribution. The larger $\nu_{k,f,0}$ and $\kappa_{k,f,0}$ are, the more k "perceptions" it takes for the parameter values of category 'k' to plateau but also the smaller the updates after each trial.

Constant delta rule. The above update equations implicitly assume that the environment is stable and therefore updated parameters keep information from all previous trials. This is the result of the generative model, which did not include a model for environmental parameter changes. Introducing expectations about environmental changes led us to consider rules with constant learning rates. We restrict ourselves here to updates for the six location parameters ($\theta_{k,V}$ $\theta_{k,A}$) of Eq 4.

We first considered a constant delta rule scaled by the evidence in favour of the selected category $p(k|s_V, s_A)$,

$$
\Delta\theta_{k,f} = A(\bar{s}_f - \theta_{k,f})\ p(k|s_V, s_A) \tag{8}
$$

As in the Bayesian case, updates accumulate without decay between trials. The main difference is that $\theta_{k,f}$ is driven more strongly by recent evidence than by past evidence, implicitly acknowledging the presence of volatility.

Hierarchical delta rule with decay. Finally we consider updates that decay with time. We reasoned that the decay should be towards parameter estimates that are more stable, which we denote by $\mu_{k,f}$. We propose a hierarchical relation, with updates in $\mu_{k,f}$ being driven by $\theta_{k,f}$, while updates in $\theta_{k,f}$ are driven by sensory evidence. At each trial all categories (k') decay toward their long-term estimates and only the perceived category (k) updates both $\mu_{k,f}$ and $\theta_{k,f}$:

$$
\begin{aligned}
\Delta\theta_{k',f} &= B(\mu_{k'} - \theta_{k',f}) \\
\Delta\theta_{k,f} &= A(\bar{s}_f - \theta_{k,f})\, p(k|s_V, s_A) \\
\Delta\mu_{k,f} &= C(\theta_{k,f} - \mu_{k,f})\, p(k|s_V, s_A)
\end{aligned} \tag{9}
$$

The first equation reflects the decay while the last two equations apply to the perceived category 'k'. 'A', 'B' and 'C' are constant parameters.

Model simulations

We simulate the experimental paradigm in Lüttke et al. (Lüttke et al. 2016), in which human participants were asked about what they heard when presented with auditory syllables or auditory syllables accompanied with a video of the corresponding speaker's lip movements. There were 6 stimulus types: three acoustic only stimuli: /aba/, /ada/ and /aga/ and three audiovisual stimuli, congruent /aba/, congruent /ada/ and McGurk stimuli, that is, acoustic /aba/ accompanied by the video of a speaker articulating /aga/. In the original experiment three different realizations of each of the 6 types were used. In our simulations we use a single realization per stimulus.

Our model simulates a single participant. As described above it proposes that syllables are encoded in terms of the expected amplitudes and variances of audiovisual features. The expected amplitudes were taken from the mean values across 10 productions from a single male speaker (Olasagasti, Bouton, and Giraud 2015), the amplitudes were then normalized by dividing by the highest value for each feature. The remaining parameters in the model, variances and sensory noise levels, were chosen so that the overall categorization results, percentage of /aba/, /ada/ and /aga/ responses to the 6 types of stimuli were qualitatively similar to those reported by Lüttke and colleagues. There was no explicit effort to fit the model, since our goal is to present a plausibility argument. What we asked was that acoustic and congruent audiovisual items were mostly correctly categorized and that McGurk stimuli were mostly categorized as /ada/, thus fulfilling the conditions of the participants chosen for the analysis in (Lüttke et al. 2016).

The model parameter values are given in Table 1.

Table 1. **Model parameters**.

$\theta_{/aba/,A} = 0.1$	$\theta_{/ada/,A} = 0.4$	$\theta_{/aga/,A} = 1$	$\sigma_{k,C} = 0.1$	$\sigma_A = 0.12$
$\theta_{/aba/,V} = 1$	$\theta_{/ada/,V} = 0.6$	$\theta_{/aga/,V} = 0.37$	$\sigma_{k,C} = 0.1$	$\sigma_V = 0.06$

The first three columns correspond to the initial location parameters of the three categories along the acoustic and visual feature dimensions. The 4th column gives the initial standard deviations of the three categories. The last column gives the values of sensory noise in the two modalities.

The stimuli presented to the modelled participant corresponded to the expected acoustic and visual features of their internal model. Thus if the internal model for /aba/ is centered at $\theta_{/aba/,A}$ for the acoustic feature and at $\theta_{/aba/,V}$ for the visual feature, those are the amplitudes chosen for the input stimuli. In other words, stimuli were tailored to the modelled participant. The six stimuli were defined by:

Acoustic /aba/	$(C_V\ C_A) = (\ \varnothing\ \ \theta_{/aba/,A})$
Acoustic /ada/	$(C_V\ C_A) = (\ \varnothing\ \ \theta_{/ada/,A})$
Acoustic /aga/	$(C_V\ C_A) = (\ \varnothing\ \ \theta_{/aga/,A})$
Congruent /aba/	$(C_V\ C_A) = (\theta_{/aba/,V}\ \theta_{/aba/,A})$
McGurk	$(C_V\ C_A) = (\theta_{/aga/,V}\ \theta_{/aba/,A})$

Congruent /aga/ $(C_V\ C_A) = (\theta_{/aga/,V}\ \theta_{/aga/,A})$

Even if the underlying parameters for a given stimulus type were the same for every trial, sensory noise created variability. The input to the model was the pair of temporal sequences $y_V(t)$ and $y_A(t)$ defined by:

$$y_V(t) = C_V\, M_V(t) + \sigma_V\, \nu(t)$$

$$y_A(t) = C_A\, M_A(t) + \sigma_A\, \nu(t)$$

The lip and 2nd formant temporal modulation profiles ($M_V(t)$ and $M_A(t)$, Fig 1B) were defined as in (Olasagasti, Bouton, and Giraud 2015). Each profile, representing the intervocalic transition between the two vowels in the "a" vocalic context, was modelled with 289 time points. The profiles are normalized so that the sum of $M^2_{A,n}$ over 'n' (number of time points) is one. The last term represents independent Gaussian noise different for the two modalities.

The six stimulus types were presented to the model in random order. At the end of the presentation, the model chose a percept based on the posterior distribution over syllable identity 'k' given the stimulus. The perceived syllable was then recalibrated by updating its defining parameters (either both mean and variance or mean alone).

We looked at acoustic only /aba/ trials and compared the proportion of /ada/ responses in two "contrasts" as in Lüttke et al. (Lüttke et al. 2016). The McGurk contrast compares the proportion of /ada/ responses when the preceding trial was a fused McGurk trial versus other stimuli (acoustic /aba/ and /aga/ and congruent /aba/ and /aga/). The "ada" contrast looks at acoustic /aba/ trials preceded by correctly categorized acoustic /ada/ versus other stimuli (acoustic /aba/ and /ada/).

All simulations were performed using custom scripts written in MATLAB (Release R2014b, The MathWorks, Inc., Natick, Massachusetts, United States). Pair-wise comparisons were performed using MATLAB's implementation of the Wilcoxon signed rank test.

While the original experiment had 27 participants and each stimulus type was presented 69 times, we run the experiment 100 times with 200 repetitions per stimulus type. The larger number of repetitions per experiment allowed a reduction in the sampling variability in the percentages of acoustic /aba/ categorized as /ada/. 95% confidence intervals were estimated by randomly sampling 27 simulations from the 100 available. For each simulation, the median difference between the proportion of acoustic /aba/ categorized as /ada/ after fused McGurk or other trials was calculated. The sampling described above was used to estimate the 95% confidence interval of the increase in proportion of /ada/ percepts after a fused McGurk trial.

Results

Simulation of recalibration

The McGurk effect offers a unique opportunity to assess cross-modality effects in adaptation. Lüttke et al. (Lüttke et al. 2016) in a re-assessment of an existing dataset selected the listeners that showed a high percentage of fused percepts for McGurk stimuli. These listeners were combining information from both modalities since the same /aba/ acoustic token was correctly categorized as /aba/ when presented alone. These participants classified acoustic only /aba/ stimuli as /ada/ more frequently if the preceding trial was a fused McGurk trial. Specifically, the percentage of acoustic /aba/ classified as /ada/ was 29% after a fused McGurk trial, versus 16% when preceded by other stimuli.

Our goal was to present a model that shows how this effect fits within an ideal observer approach that uses an internal generative model to interpret sensory input, and which continuously adapts the generative model to better match the incoming input. The model characterizes the sensory input by a visual feature, the amplitude of lip closure in the transition between the two vowels,

which we denote by 'V'; and an acoustic feature, the amplitude of the 2nd formant transition, which we denote by 'A'.

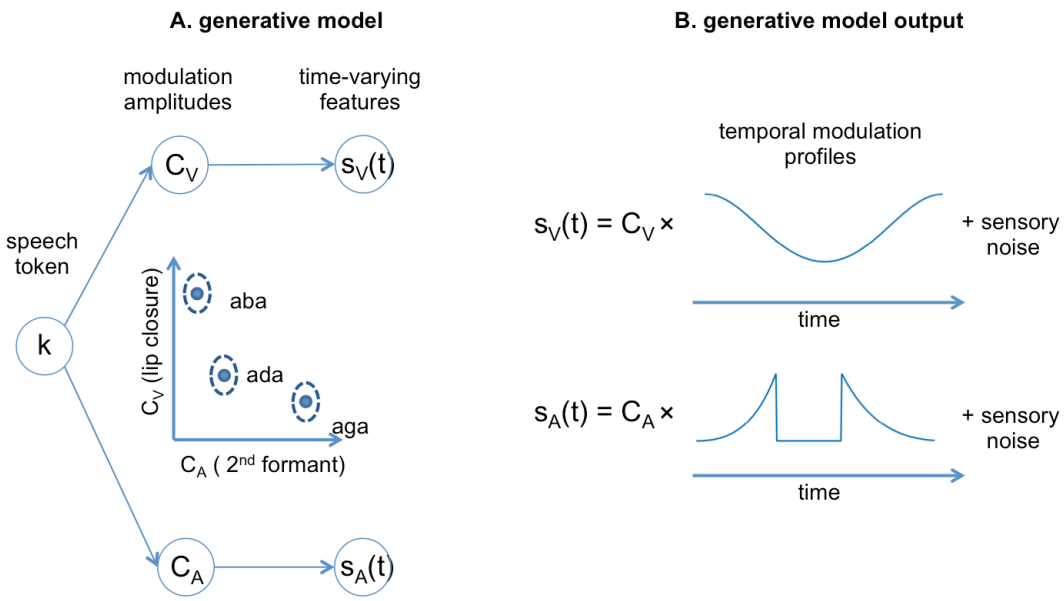


Fig 1. **Schematics of the generative model**. A. The schematic on the left shows the generative model. In a given trial, a speech token "k" determines the amplitudes of degree of lip closure (C_V) and magnitude of second formant deflection (C_A) by sampling from the appropriate Gaussian distribution. The distributions corresponding to each "k" are represented in the two-dimensional feature space in the middle. These values determine the amplitude of the time varying audiovisual features. B**.** The right panels show in more detail how the time varying output of the generative model is determined given a pair (C_V, C_A). The temporal profiles are fixed in the present model and are common for the three tokens, which are assumed to change only in their magnitude. The model also includes the "injection" of sensory noise. The lip modulation profile shows the lip movement in the transition between the two vowels. Similarly, the 2nd formant transition profile shows the transition in the second formant between the two vowels.

The model, described in detail in the methods section, reproduces the average performance across participants in the task; we chose parameters that elicit a very high frequency of McGurk percepts (Fig 2B, middle panel of the bottom row) and assumed that listeners were always integrating the two sensory streams.

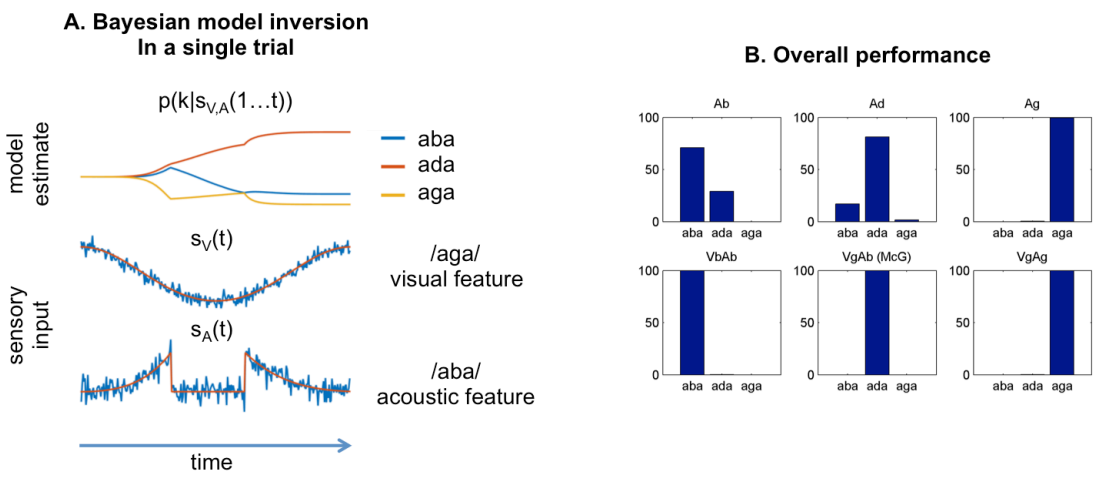


Fig 2. **Model's overall performance**. A: Stimuli are presented to the simulated participant, which inverts the generative model and gathers evidence for each of the three possible outcomes. The example corresponds to a McGurk trial (acoustic /aba/, visual /aga/). The model's posterior is maximum for /ada/. B: Simulation of the Lüttke et al. (2016) experiment. Model classification across all trials. Each subpanel shows the percentage of /aba/, /ada/ and /aga/ percepts corresponding to each of the six conditions (Ab: acoustic only /aba/; Ad: acoustic only /ada/; Ag: acoustic only /aga/; VbAb: congruent audivisual /aba/; VgAb incongruent McGurk stimuli with visual /aga/ and acoustic /aba/; VgAg: congruent /aga/). Congruent and acoustic only stimuli are categorized with a high degree of accuracy and McGurk trials are consistently fused, that is, perceived as /ada/. We reproduce the experimental paradigm consisting of six types of stimuli presented in pseudo-random order; three non-ambiguous acoustic only tokens: /aba/, /ada/ and /aga/, and three audiovisual stimuli: congruent /aba/, incongruent visual /aga/ with acoustic /aba/ (McGurk stimuli), and congruent /ada/.

In both acoustic and audiovisual trials listeners were asked to report the perceived acoustic stimulus in a three alternative forced choice task. To simulate the listener's choice, we calculated the probability of the acoustic token given the stimulus. In our notation, $p(k_A|s_V,s_A)$ for audiovisual stimuli and $p(k_A|s_A)$ for unimodal acoustic stimuli. The percept at a single trial was determined by choosing the category that maximized the posterior. Fig 2A shows the model's response to a McGurk stimulus; at the end of the trial the most likely interpretation corresponds to the syllable /ada/.

To model the recalibration after each trial, we changed the generative model's location parameters associated with the perceived category. This was done for the acoustic and visual parameters after an audio-visual trial, and for the acoustic parameter after an acoustic trial. We assumed that this happens for every trial and is part of the normal perceptual process. The changes are thus driven by any residual sensory prediction error.

When listeners consistently reported the fused percept /ada/ when confronted with a video of /aga/ and the sound of /aba/, the presence of the visual stream modified the acoustic percept from /aba/ to /ada/. Given that the acoustic input did not correspond to the one that was most expected from the perceived token, there was a systematic residual sensory prediction error. Since this residual prediction error was used as a signal to drive the model's adaptation, the /ada/ representation moved towards the McGurk stimulus parameters after a fused percept (Fig 3A).

In the model, the residual prediction error occurs in the transformation from token identity to expected modulation of the acoustic feature (C_A). Sensory evidence drives C_A towards the experimentally presented value: /aba/ for McGurk stimuli. Thus, when the percept is /ada/, there is a mismatch between the top-down prediction as determined by the top-down component $p(C|k)$ which drives C_A towards $\theta_{/ada/,k}$, and the actual value C_A(/aba/) determined by the bottom-up component. This is evident in the following expression,

$$C_{A,k} = \frac{\bar{\sigma}_A^2}{\bar{\sigma}_A^2 + \sigma_{C_A,k}^2}\theta_{A,k} + \frac{\sigma_{C_A,k}^2}{\bar{\sigma}_A^2 + \sigma_{C_A,k}^2}\bar{s}_A \qquad (10)$$

with the first term reflecting the prior expectation for category 'k' and the second reflecting the sensory evidence accumulated during the trial (see Eq 6).

The expression can be rewritten to make the prediction error explicit.

$$C_{A,k} = \theta_{A,k} + \frac{\sigma_{C_A,k}^2}{\bar{\sigma}_A^2 + \sigma_{C_A,k}^2}(\bar{s}_A - \theta_{A,k}) \qquad (11)$$

This expression highlights the prediction from the category (first term) and the residual prediction error in the second term.

To reduce the residual prediction error we consider that the participant recalibrates its generative model, that is, that it will change $\theta_{k,V}$ and $\theta_{k,A}$ driving them towards s_V and s_A respectively. If the stimuli are chosen with parameters "adapted" to the listener, as we do, $s_V \sim \theta_{stim,V}$ and $s_A \sim \theta_{stim,A}$.

We now present the results from running the experiment on 100 simulated listeners. A simulated listener is shown in Fig 2B. Each panel gives the histograms of responses (/aba/, /ada/ or /aga/) for each of the six types of stimulus. As in Lüttke et al. (Lüttke et al. 2016), accuracy for both acoustic and congruent audio-visual stimuli is high. In addition, there is a high frequency of fused percepts. We report the results obtained with different update rules; one derived from a Bayesian model as in (Kleinschmidt and Jaeger 2015), which assumes a stable environment, and two empirically motivated.

As a control we also simulated the experiment with no parameter updates (Fig 4A). The percentage of acoustic /aba/ categorized as /ada/ was 10% and the 95% CI for the difference between trials preceded or not by a fused McGurk stimulus was [-2.86 2.23]%.

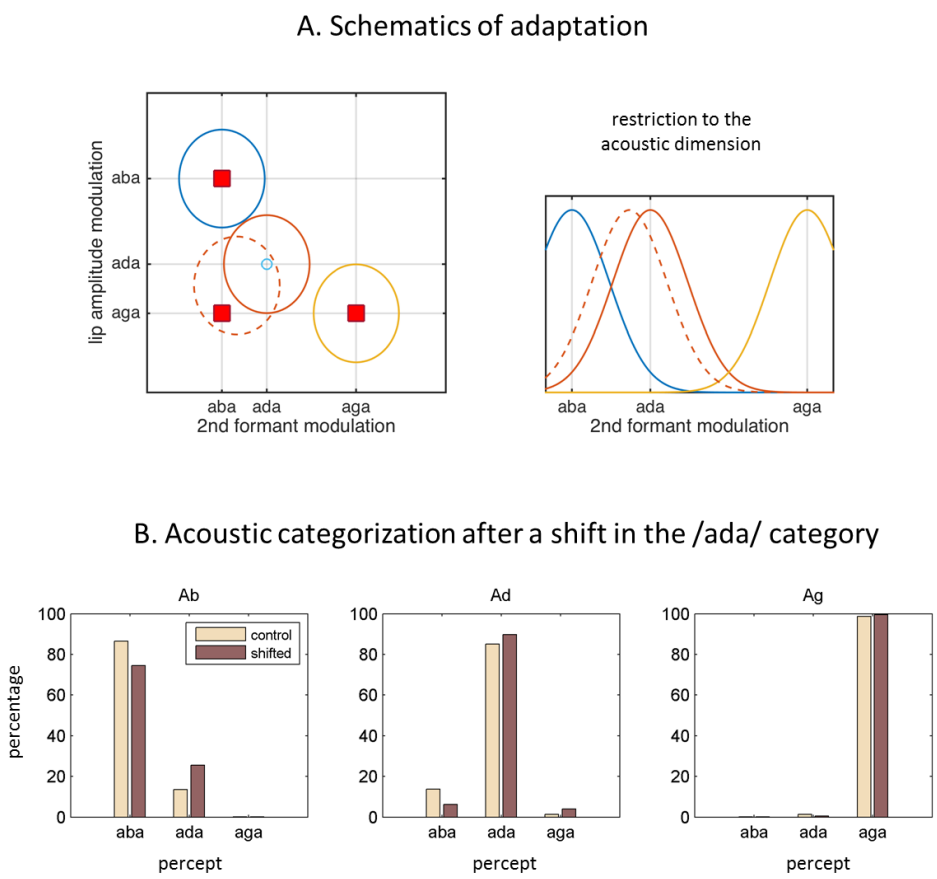


Fig 3. **Internal model adaptation**. A: Speech tokens are represented in a multimodal feature space here represented by two dimensions. Each ellipse stands for the internal representation of each congruent category ('aba' in blue, 'ada' in red, 'aga' in yellow). The red squares show the location of the audiovisual stimuli in the 2D feature space. They represent congruent /aba/ (top left), congruent /aga/ (bottom right), and McGurk stimuli (bottom left). When McGurk stimuli are repeatedly perceived as /ada/, the /ada/ representation (in solid red) is modified in such a way that it "moves" (dashed red) towards the presented McGurk stimulus (visual /aga/ with acoustic /aba/) and therefore should affect the processing of subsequent sensory input. The right panel shows the restriction to the acoustic dimension, illustrating how the acoustic representation for /ada/ has shifted towards that of /aba/. B: The effects of the shift in the internal representation on the categorization of the purely acoustic /aba/ (Ab), /ada/ (Ad) and /aga/ (Ag) sounds.

Each panel shows the percentage of /aba/, /ada/ and /aga/ percepts for the "control" representations (solid lines) and the representations with the recalibrated /ada/ (dashed line). As in Lüttke et al. (2016), the biggest effect is observed in the categorization of the /aba/ sounds.

Bayesian updating

We first considered the same kind of update rule used by Kleinschmidt and Jaeger (Kleinschmidt and Jaeger 2015) to model changes in speech sound categorization after exposure to adapting stimuli. After each trial the generative model can update the parameters by considering the probability of the parameters given the sensory input and the categorization $p(\theta|k\ s_V\ s_A)$, leading to sequential Bayesian updating (Eq 7 in the methods section; we run the simulations with $\kappa_{k,f,0} = 3$ and $v_{k,f,0} = 10$). The percentage of /ada/ responses to acoustic only /aba/ stimuli was 17% both after fused McGurk and after all other control conditions. The 95% confidence interval for the difference in medians between the two conditions was [-2.35, 3.08]%, thus failing to reproduce the effect of interest. This might not be surprising; the Bayesian update rules have the form of a delta rule with a decreasing learning rate. As a consequence, the size of the changes in the categories diminishes as the experiment progresses and all stimuli experience the same internal model. The updates did lead to observable effects; there was an overall increase of /ada/ responses to acoustic /aba/ (17% vs. the 10% of the control experiment without parameter updates). The resulting changes in model parameters would lead to an after-effect, that is, the point of subjective equivalence in an /aba/-/ada/ acoustic continuum would be shifted in the direction of /aba/.

Constant delta rule

The Bayesian update rule used above assumes that the parameters are constant in time and that therefore all samples have equal value, whether they are old or recent. This leads to a delta rule with a learning rate that tends to zero. We therefore considered a rule with a constant learning rate, which allows updates of similar magnitude over the whole experiment. The model's expected modulation for the perceived category was recalibrated according to:

$\Delta\theta_{k,f} = 0.1\ (s_f - \theta_k)\ p(k|s_A,s_V)$ only for k = percept

As Fig 4B shows, the percentage of acoustic /aba/ categorized as /ada/ was not significantly higher when the preceding trial was a fused McGurk trial compared with any other stimulus (34% vs. 33%, median difference at 95% CI [-2.7 7.1]%). Note that in this case too, the proportion of /ada/ responses for acoustic /aba/ inputs is increased compared with simulations run without adaptation (Fig 4A).

Although the learning rate is constant, which means that recalibration magnitude does not decrease during the experiment, recalibration does not decay across trials.

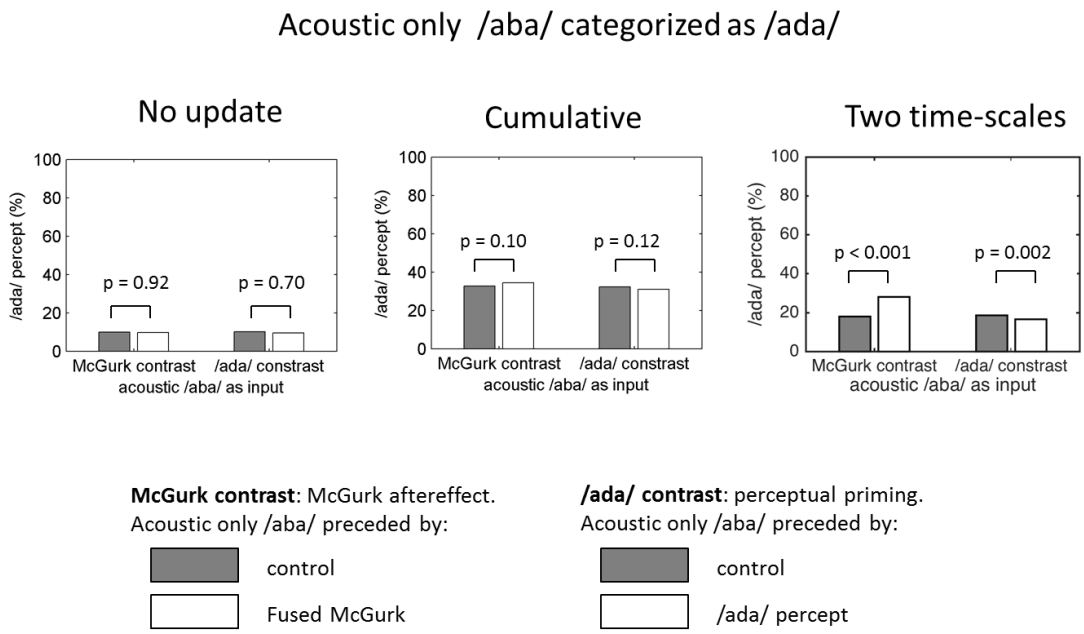


Fig 4. **Cumulative and transient update rules**. /ada/ percepts in response to acoustic /aba/ stimulation. We show two contrasts. The McGurk contrast compares the percentage of /ada/ responses when acoustic /aba/ is preceded by control stimuli (acoustic /aba/ and /aga/, congruent /aba/ and /aga/) versus fused McGurk trials. The /ada/ contrast refers to acoustic /aba/ preceded by control stimuli (acoustic /aba/ and /aga/) versus acoustic trials correctly perceived as /ada/. The three panels show the result of running the same perceptual model supplemented by three different update rules. Left, control, no update. Middle: constant delta rule. Right: update rule assuming that modification occurs at two time scales. The constant delta rule (middle) leads to changes in internal representations that are reflected in the overall increase of /ada/ percepts (with respect to the control, no update model on the left) however, it does not translate into effects on the previous trial. The model assuming two time scales does reproduce the effect of the fused McGurk on the next trial (McGurk contrast).

Hierarchical Updates with intrinsic decay

We tested an alternative update rule based on the idea that it should reflect how changes occur in the environment. In this respect, we propose that one can characterize changes as occurring hierarchically; with just two levels at a first approximation. These would correspond to keeping "running averages" over different time scales, to be sensitive to fast changes without forgetting about longer-lasting trends. Furthermore, recalibration is probably corrupted by interference or noise over time.

We consider two sets of hierarchically related variables $\theta_{k,f}$ (fast) and $\mu_{k,f}$ (slow). The faster decaying one, $\theta_{k,f}$, is driven by both sensory prediction error and the more slowly changing variable, $\mu_{k,f}$ (more details can be found in Materials and Methods). The more slowly changing and decaying variable, $\mu_{k,f}$, keeps a representation based on a longer term "average" over sensory evidence. In the limit, $\mu_{k,f}$ is constant; and this is what we consider here for illustrative purposes. Thus the update rules include a "jump" in the fast variable due to the sensory prediction error in the perceived category plus a decay term toward the slower variable for every category. The "jump" corresponds to the traditional update after an observation, the decay reflects the transient character of this update.

The results in Fig 4C were run with the following update rules:

$\Delta\theta_{k,f} = 0.1\ (\mu_{k,f} - \theta_{k,f})\ p(k|s_A,s_V)$ for all k

$\Delta\theta_{k,f} = 0.3\ (s_f - \theta_k)\ p(k|s_A,s_V)$ only for k = percept

$\Delta\mu_{k,f} = 0$ for all k

where subscript f indexes the feature (f = A for acoustic, f = V for visual). All categories decay toward the corresponding long-term stable values ($\mu_{k,A}$, $\mu_{k,V}$) in the inter-trial interval.

The percentage of acoustic /aba/ categorized as /ada/ after control trials was 18% vs. 28% after fused McGurk, the median difference being 95% CI: [5.8 13.0]%. Therefore two effects can be observed; the overall increase in acoustic /aba/ categorized as /ada/ and the rapid recalibration effect reflected in the specific increase observed when acoustic /aba/ was preceded by a fused McGurk trial.

Discussion

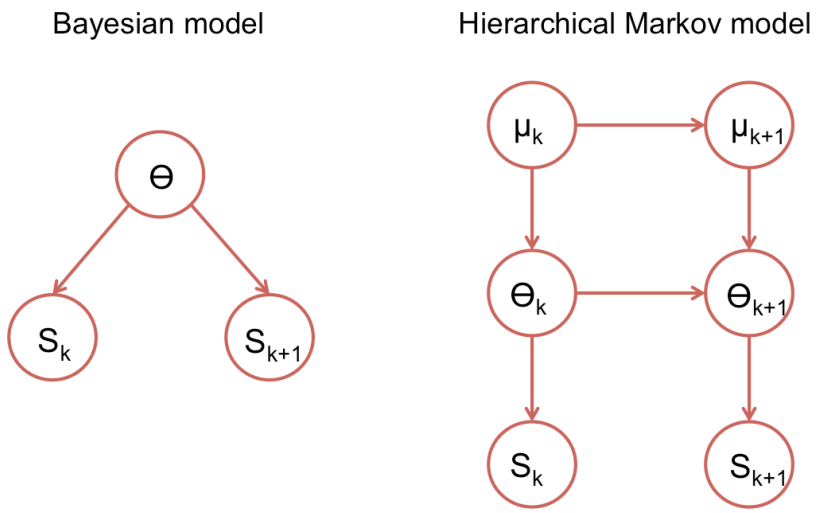
Speech sound categories are sensitive to the most recently presented stimuli. In this modelling work we provide a possible account for rapid recalibration of speech sound categories in response to a changing sensory environment. Our account suggests that recalibration occurs at least at two levels characterized by two different decaying timescales. We propose a fast decaying/short-term level (on the order of seconds) that weighs recent versus past sensory evidence more strongly than a slowly decaying/long-term level that keeps a longer record of past sensory evidence. These elements are important because speech sounds have both variable and stable components; we want to adapt to a new speaker without forgetting our past experience with other speakers.

Our model was motivated by the experimental results by Lüttke et al. (Lüttke et al. 2016) showing that the categorization of /aba/ sounds differed depending on whether the previous trial was a fused McGurk trial versus any of the other five speech stimulus types presented in the course of their experiment. Since this effect was distinct from other well-documented across-trial dependency effects such as perceptual priming or selective adaptation, we sought to model the effect solely considering changes in perceived speech categories without external feedback. Like previous approaches, ours builds on the idea that the brain achieves perception by inverting a generative model (e.g. (Rao and Ballard 1999; Knill and Pouget 2004; Friston 2005)) and by continuously monitoring the performance of this model to adapt it to the current stimulus landscape. One way the brain can change the model without external feedback is by using the perceptual outcome as teaching signal. The latter tries to alter the model so that the perceptual outcome better aligns with the sensory features leading to it (e.g. (Kleinschmidt and Jaeger 2015)). Humans seem to use their own categorical perceptual decisions as if they were veridical to inform a sensory identification task (Luu and Stocker 2018).

Recalibration of speech sound categories has been described within an “ideal adapter” framework (Kleinschmidt and Jaeger 2015). In this framework, speech sound categories are subject to trial-by-trial changes well described by a Bayesian approach that implicitly assumes that sound categories are stable within an experimental session, and hence uses update/learning rules that give the same weight to recent and past evidence. This makes sense in a stable environment where information remains equally relevant independent of its recency. While this adaptation rule successfully described the dynamics of adaptation in sublexical speech categories in experiments with blocks of constant repeated stimuli (Bertelson, Vroomen, and De Gelder 2003; Vroomen et al. 2007), it failed here to account for the specific transient effect in Lüttke et al. (Lüttke et al. 2015).

Yet, we were able to successfully reproduce the results of (Lüttke et al. 2016) by proposing empirically motivated update equations that include two hierarchically related variables with different learning and decaying timescales; a fast changing variable driven by current sensory prediction errors, and a more slowly changing one that acts as a longer-span buffer (Fig 5,

equations (8) and (9)). As a result, the variable controlling the model expectations decays toward the longer-term variable after a transient change driven by the current sensory prediction error. It turns out that the longer term variable encodes the cumulative effects observed in the typical adaptation experiments (e.g., (Bertelson, Vroomen, and De Gelder 2003)). Thus, after a fused McGurk trial, the acoustic representation of the /ada/ category was shifted towards that of acoustic /aba/ (Fig 3A), which increased the chances of an acoustic /aba/ to be categorized as /ada/ (Fig 3B). Since the shift decays over time, the increase in /ada/ percepts for acoustic /aba/ was most prominent when the acoustic /aba/ trial immediately followed the fused McGurk trial. On the other hand, since the decay goes towards a slower changing "version" of the /ada/ category, the model could also show a more persistent accumulated adaptation as in Kleinschmidt and Jaeger (Kleinschmidt and Jaeger 2015).



$$\hat{\Theta}_{k+1} = \hat{\Theta}_k + \alpha(s_k - \hat{\Theta}_k) \qquad \hat{\Theta}_{k+1} = \hat{\Theta}_k + \alpha(s_k - \hat{\Theta}_k) + \beta(\hat{\mu}_k - \hat{\Theta}_k)$$

$$\alpha \rightarrow 0 \qquad \hat{\mu}_{k+1} = \hat{\mu}_k + \lambda(\hat{\Theta}_{k+1} - \hat{\mu}_k)$$

Fig 5. **Statistical models underlying the two classes of update rules used in the paper**. Here Θ stands for the model parameters that determine the speech categories used by the perceptual model (See Fig 3) and 'k' for the trial index. On the standard Bayesian approach (left), model parameters are considered constant in time leading to update rules that give the same weight to all prediction errors, which in turn leads to a "learning rate" α_k that becomes smaller with the number of trials. On the right we show a hierarchical Markov model implementation that would lead to the kind of update rules that we introduced empirically to accommodate the rapid recalibration effect. This alternative view implicitly assumes that model parameters can change in time and therefore lead to update rules with learning rates (α, β and λ) that settle to non-zero constant values.

In this work, we focused on the recalibration of the acoustic features associated with the speech categories. However, given that lipreading can also be recalibrated by audiovisual input (Baart and Vroomen 2010), we would predict a corresponding effect on lipreading, namely that a purely visual /aga/ would be more readily categorized as /ada/ immediately after a fused McGurk trial. In summary, we propose that during audiovisual speech integration, and in McGurk stimuli in particular, the brain tries to find the most parsimonious and general explanation for the input even at the expense of residual prediction errors at the sensory periphery. Different participants may value differently this parsimony/accuracy trade-off; we hypothesize that those who consistently fuse the two streams change their templates for the /ada/ category, that is for the perceived

category, rather than for how the acoustic and visual features are processed in the periphery. The change in /ada/ effectively leads to category boundary shifts in acoustic continua, boundary shifts that are transient and will only add-up after consistent repetitions. Importantly, these effects are distinct from perceptual bias and specific adaption.

Activity in auditory cortex. Lüttke et al. (Lüttke et al. 2016) also noted that the behavioural effect was accompanied by distinct activity in auditory cortex. When an acoustic only /aba/ was presented after a fused McGurk trial, fMRI activity in auditory cortex was more frequently classified as /ada/ by a learning algorithm trained to distinguish correctly identified acoustic /aba/, /ada/ and /aga/. We have argued above that the intermediate level in our perceptual model could be interpreted as the representation of acoustic features in auditory cortex and visual features in extrastriate visual cortex. The acoustic feature in the model is the amplitude of the 2nd formant transition: C_A. As the perceptual process progresses within the trial, C_A is updated by moment-to-moment bottom-up prediction errors and by updated top-down predictions dominated by the /ada/ category in trials that will subsequently be categorized as /ada/ by the simulated participants. Since an /ada/ percept for acoustic /aba/ occurs more frequently after a fused McGurk trial, C_A will correspondingly be closer to C_A(ada) than on trials preceded by the other types of stimuli, which lead to a lower proportion of /ada/ percepts. Therefore, our generative model, just as shown by the fMRI data, would predict activity in auditory cortex closer to /ada/ during the acoustic /aba/ after the fused McGurk trials, owing to the top-down input from the speech category level in the course of the within trial perceptual process.

Revised "ideal adapter". While the "ideal adapter" from Kleinschmidt and Jaeger (Kleinschmidt and Jaeger 2015) focused on cumulative recalibration effects, our results suggest that shorter-lived effects are also behaviorally relevant. While the ideal adapter could be formalized as incremental optimal Bayesian inference in a non-volatile environment (Fig 5A), our empirically motivated update rule could possibly be cast in a normative framework by explicitly accounting for the volatility in the environment. True changes in the environment are sometimes referred to as 'volatility' to distinguish them from the trial-by-trial variability observable in a fixed environment. A hierarchical Markov model as the one shown in the right panel of Fig 5 could translate into hierarchical update equations of the form of the empirically proposed ones (Wilson, Nassar, and Gold 2013). This is because in a volatile environment current information is more relevant than past information, which leads to the notion of "optimal forgetting". If the level of volatility is stable, one obtains a delta rule with a learning rate that converges onto a constant, with higher learning rates (faster "forgetting") related to higher volatility levels. In Fig 5B it is assumed that the higher level (μ) has lower volatility than the intermediate level (θ). Therefore, our model combines volatility with hierarchy. This combination departs from models used to explain decision making in changing environments (Behrens et al. 2007; Summerfield, Behrens, and Koechlin 2011; C. D. Mathys et al. 2014), which are not hierarchical, and focus on the nontrivial task of inferring the volatility of the environment. The above studies provide evidence that human participants do adapt their learning rates when volatility is changed. Human performance in these tasks could be modelled without the need to keep representations over several time scales. It is important to note that in these tasks participants need to keep track of changes in arbitrary cue-reward associations or in arbitrarily defined sensory categories (Summerfield, Behrens, and Koechlin 2011). Our work refers to the adaptation of sublexical speech categories, that is, of already existing, learned and physiologically relevant categories. The brain, therefore should not simply keep an estimate that changes flexibly to the local temporal environment, but probably benefits from also keeping a long-term estimate as an empirical "prior".

Interestingly, hierarchically related variables with increasing time scales have been shown in a modelling study to increase the capacity of memory systems and improve the stability of synaptic modifications (Benna and Fusi 2016); and a model of synapses with a cascade of metastable states with increasing stability can flexibly learn under uncertainty (Iigaya 2016). The latter also modelled the intrinsic decay of synaptic modifications. The decay was faster for the more labile

memory states, which are also those memory states that are more sensitive to new evidence. The deeper, more stable memory states showed slower decay. This maps nicely to our proposal.

Temporal hierarchies might also reflect the hierarchical nature of perception, in which higher level representations are increasingly abstract. If we see perception as a tower of abstractions, adaptation might work at every level of the hierarchy, with more abstract categories integrating update information within increasingly longer time windows, and therefore being increasingly stable.

Our modelling data are relevant in the domain of continuous speech processing, in particular to account for auditory processing anomalies in dyslexia. Evidence from a two-tone frequency discrimination task suggests that participants' choices are driven not only by the tones presented at a given trial, but also by the recent history of tone frequencies in the experiment, with recent tones having more weight than earlier ones (Jaffe-Dax, Frenkel, and Ahissar 2017). It turns out that, when compared with controls, dyslexics show a decreased reliance on temporally distant tones, suggesting a shorter time constant (Jaffe-Dax et al., 2017). Translating this result to the current model, we could hypothesize that in dyslexia the long-term component (μ in Fig 5B) has either a shorter time span or it is coupled to the lower representation with a lower weight. In both cases, we expect a deficit in building long-term stable speech category representations since they would be overly driven by the current context.

Conclusion

We have presented a revised "ideal adapter" model for speech sound recalibration that has both transient and cumulative components organized hierarchically.

This new model allows us to interpret Lüttke et al.'s finding (2016) as evidence for a hierarchy of processes in the recalibration of speech categories. Our model further highlights that after experiencing the McGurk effect, it is not the acoustic features related to the sensory input (/aba/) that are modified, but the higher-level syllabic representation related to the percept (/ada/). This suggests that part of the changes observed in the activity of sensory cortices that encode specific acoustic cues are not locally generated but are a reflection of an interaction between bottom-up peripheral sensory inputs and top-down expectations from regions where categorical perception takes place.

As far as speech perception is concerned, our results show that it might be fruitful to consider the perception of speech in an ever changing acoustic environment (new speakers, languages, presence of noise etc…) as the inversion of a continuously monitored internal generative model adapted at different temporal scales closely reflecting its hierarchical structure.

References

Baart, Martijn, and Jean Vroomen. 2010. "Do You See What You Are Hearing? Cross-Modal Effects of Speech Sounds on Lipreading." *Neuroscience Letters* 471 (2): 100–103. https://doi.org/10.1016/j.neulet.2010.01.019.

Behrens, Timothy E J, Mark W Woolrich, Mark E Walton, and Matthew F S Rushworth. 2007. "Learning the Value of Information in an Uncertain World." *Nature Neuroscience* 10 (9): 1214–21. https://doi.org/10.1038/nn1954.

Benna, Marcus K, and Stefano Fusi. 2016. "Computational Principles of Synaptic Memory Consolidation." *Nature Neuroscience* 19 (12): 1697–1706. https://doi.org/10.1038/nn.4401.

Bertelson, Paul, Jean Vroomen, and Béatrice De Gelder. 2003. "Visual Recalibration of Auditory Speech Identification: A McGurk Aftereffect." *Psychological Science* 14 (6): 592–97. http://www.ncbi.nlm.nih.gov/pubmed/14629691.

Campbell, R. 2008. "The Processing of Audio-Visual Speech: Empirical and Neural Bases." *Philosophical Transactions of the Royal Society B: Biological Sciences* 363 (1493): 1001–

10. https://doi.org/10.1098/rstb.2007.2155.

Friston, Karl J. 2005. “A Theory of Cortical Responses.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 360 (1456): 815–36. https://doi.org/10.1098/rstb.2005.1622.

Gelman, Andrew, John Carlin, Hal Stern, and Donald Rubin. 2003. *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC. citeulike-article-id:105949.

Idemaru, Kaori, and Lori L Holt. 2014. “Specificity of Dimension-Based Statistical Learning in Word Recognition.” *Journal of Experimental Psychology. Human Perception and Performance* 40 (3): 1009–21. https://doi.org/10.1037/a0035269.

Iigaya, Kiyohito. 2016. “Adaptive Learning and Decision-Making under Uncertainty by Metaplastic Synapses Guided by a Surprise Detection System.” *ELife* 5. https://doi.org/10.7554/eLife.18073.

Jaffe-Dax, Sagi, Or Frenkel, and Merav Ahissar. 2017. “Dyslexics’ Faster Decay of Implicit Memory for Sounds and Words Is Manifested in Their Shorter Neural Adaptation.” *ELife* 6 (January). https://doi.org/10.7554/eLife.20557.

Kleinschmidt, Dave F, and T Florian Jaeger. 2015. “Robust Speech Perception: Recognize the Familiar, Generalize to the Similar, and Adapt to the Novel.” *Psychological Review* 122 (2): 148–203. https://doi.org/10.1037/a0038695.

Knill, David C, and Alexandre Pouget. 2004. “The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation.” *Trends in Neurosciences* 27 (12): 712–19. https://doi.org/10.1016/j.tins.2004.10.007.

LIBERMAN, A M, K S HARRIS, H S HOFFMAN, and B C GRIFFITH. 1957. “The Discrimination of Speech Sounds within and across Phoneme Boundaries.” *Journal of Experimental Psychology* 54 (5): 358–68. http://www.ncbi.nlm.nih.gov/pubmed/13481283.

Lüttke, Claudia S., Matthias Ekman, Marcel A. J. van Gerven, and Floris P. de Lange. 2016. “McGurk Illusion Recalibrates Subsequent Auditory Perception.” *Scientific Reports* 6 (September): 32891. https://doi.org/10.1038/srep32891.

Lüttke, Claudia S, Matthias Ekman, Marcel A J van Gerven, and Floris P de Lange. 2015. “Preference for Audiovisual Speech Congruency in Superior Temporal Cortex.” *Journal of Cognitive Neuroscience*, 1–7. https://doi.org/10.1162/jocn_a_00874.

Luu, Long, and Alan A Stocker. 2018. “Post-Decision Biases Reveal a Self-Consistency Principle in Perceptual Inference.” *ELife* 7 (May). https://doi.org/10.7554/eLife.33334.

Magnotti, John F., and Michael S. Beauchamp. 2017. “A Causal Inference Model Explains Perception of the McGurk Effect and Other Incongruent Audiovisual Speech.” Edited by Samuel J. Gershman. *PLOS Computational Biology* 13 (2): e1005229. https://doi.org/10.1371/journal.pcbi.1005229.

Mathys, Christoph D, Ekaterina I Lomakina, Jean Daunizeau, Sandra Iglesias, Kay H Brodersen, Karl J Friston, and Klaas E Stephan. 2014. “Uncertainty in Perception and the Hierarchical Gaussian Filter.” *Frontiers in Human Neuroscience* 8 (November). https://doi.org/10.3389/fnhum.2014.00825.

Mathys, Christoph, Jean Daunizeau, Karl J Friston, and Klaas E Stephan. 2011. “A Bayesian Foundation for Individual Learning under Uncertainty.” *Frontiers in Human Neuroscience* 5: 39. https://doi.org/10.3389/fnhum.2011.00039.

McQueen, James M, Anne Cutler, and Dennis Norris. 2006. “Phonological Abstraction in the Mental Lexicon.” *Cognitive Science* 30 (6): 1113–26. https://doi.org/10.1207/s15516709cog0000_79.

Olasagasti, Itsaso, Sophie Bouton, and Anne-Lise Giraud. 2015. “Prediction across Sensory Modalities: A Neurocomputational Model of the McGurk Effect.” *Cortex* 68 (July): 61–75. https://doi.org/10.1016/j.cortex.2015.04.008.

Rao, R P, and Dana H Ballard. 1999. “Predictive Coding in the Visual Cortex: A Functional

Interpretation of Some Extra-Classical Receptive-Field Effects." *Nature Neuroscience* 2 (1): 79–87. https://doi.org/10.1038/4580.

Reinisch, Eva, David R Wozny, Holger Mitterer, and Lori L Holt. 2014. "Phonetic Category Recalibration: What Are the Categories?" *Journal of Phonetics* 45 (July): 91–105. https://doi.org/10.1016/j.wocn.2014.04.002.

Summerfield, Christopher, Timothy E J Behrens, and Etienne Koechlin. 2011. "Perceptual Classification in a Rapidly Changing Environment." *Neuron* 71 (4): 725–36. https://doi.org/10.1016/j.neuron.2011.06.022.

Varnet, Léo, Kenneth Knoblauch, Fanny Meunier, and Michel Hoen. 2013. "Using Auditory Classification Images for the Identification of Fine Acoustic Cues Used in Speech Perception." *Frontiers in Human Neuroscience* 7 (January): 865. https://doi.org/10.3389/fnhum.2013.00865.

Vroomen, Jean, Sabine van Linden, Béatrice de Gelder, and Paul Bertelson. 2007. "Visual Recalibration and Selective Adaptation in Auditory-Visual Speech Perception: Contrasting Build-up Courses." *Neuropsychologia* 45 (3): 572–77. https://doi.org/10.1016/j.neuropsychologia.2006.01.031.

Wilson, Robert C, Matthew R Nassar, and Joshua I Gold. 2013. "A Mixture of Delta-Rules Approximation to Bayesian Inference in Change-Point Problems." *PLoS Computational Biology* 9 (7): e1003150. https://doi.org/10.1371/journal.pcbi.1003150.