

Bayesian gates: a probabilistic modeling tool for temporal segmentation of sensory streams into sequences of perceptual accumulators

Mamady Nabé, Jean-Luc Schwartz, Julien Diard

▶ **To cite this version:**

Mamady Nabé, Jean-Luc Schwartz, Julien Diard. Bayesian gates: a probabilistic modeling tool for temporal segmentation of sensory streams into sequences of perceptual accumulators. CogSci 2022 - 44th Annual Conference of the Cognitive Science Society, Jul 2022, Toronto, Canada. hal-03747788

HAL Id: hal-03747788
https://hal.science/hal-03747788

Submitted on 8 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Bayesian gates: a probabilistic modeling tool for temporal segmentation of sensory streams into sequences of perceptual accumulators

Mamady Nabé (mamady.nabe@univ-grenoble-alpes.fr)
Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 38000 Grenoble, France

Jean-Luc Schwartz (jean-luc.schwartz@gipsa-lab.grenoble-inp.fr)
Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

Julien Diard (julien.diard@univ-grenoble-alpes.fr)
Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 38000 Grenoble, France

Abstract

To explain how perception processes are performed, understanding how continuous sensory streams are temporally segmented into discrete units is central. This is particularly the case in speech perception where temporal segmentation is key for identifying linguistic units contained between consecutive events in time. We propose an original probabilistic construct, that we call "Bayesian gates", to segment temporally continuous streams of sensory stimuli into sequences of decoders. We first define Bayesian gates mathematically and describe their properties. We then illustrate their behavior in the context of a model of word recognition in speech perception. We show that, based on an event detection module, they sequentially parse the acoustic stimulus, so that each syllable decoder only processes a segment of the sensory signal.

Keywords: probabilistic modeling; temporal segmentation; perceptual accumulation; syllabic onset; speech perception

Introduction

Sensory processes appear sometimes as being "mostly continuous", in the sense that, when a physical characteristic of the input signal is measured by the system, the relation between the external physical measurement and the internal perception may be described by a smoothly varying transfer function. Consider for instance loudness perception in the auditory pathway, which does not feature evident discontinuities (Stevens, 1955; Zwicker & Sharf, 1965; Schlittenlacher & Ellermeier, 2021).

Speech perception appears to contrast with this. Speech signals are produced from discrete linguistics units ordered sequentially through time, at the sentence, phrase, word, syllabic and phonemic levels; although this last level is debatable (Cutler, Mehler, Norris, & Segui, 1987; Lotto & Holt, 2000). Therefore, a major component of speech perception concerns processing a continuous speech signal into discrete units. This contains two "discretization issues". The first one is to map variable realizations in the acoustic domain to speech units that most likely yielded them: this is a recognition (categorization) issue. The second one is to identify, in the speech signal, temporal intervals and their boundaries, that correspond to each linguistic unit in their realization order: this is a temporal segmentation issue. These two discretization issues are, of course, not treated sequentially, nor independently, and appear instead largely dependent on each other.

A classical class of computational, probabilistic models for speech processing relies on Hidden Markov Models (HMMs) (Rabiner, 1989; Gales & Young, 2008). In such models, a state variable represents "decoding steps", which may or may not align with and correspond to linguistic units. To parse a speech signal, HMMs start in a given initial state; the probability distribution over states, or over state sequences, then evolves over time as the sensory signal is processed. After termination, the probability distribution over state sequences corresponds to the most likely interpretation of the input signal in terms of sequences of linguistic units. However, in such models, the temporal segmentation of the input is not explicit or directly interpretable. In other words, HMMs do not necessarily yield a one-to-one mapping between linguistic units and internal states: on the one hand, the HMM state can vary during the processing of a single phoneme (for instance, during a vowel, to separate the initial portion, "contaminated by the previous consonant", from later portions, that could be "contaminated by the upcoming consonant"); on the other hand, a single HMM state can represent a portion of the acoustic input that straddles a boundary between linguistic units.

This is also true in most connectionist models. In the classical TRACE model, for instance, temporal segmentation is not explicit, nor necessarily aligned with linguistic units (McClelland & Elman, 1986). In recent models based on deep learning, latent representations are learned from data, once again with no guarantee that they would align with linguistic units, neither concerning their acoustic content, nor concerning their temporal properties and boundaries (Girin et al., 2021).

This contrasts with a new generation of speech perception models informed by neuroanatomical constraints, in which temporal segmentation of the input is a central, explicit concern. For instance, see the TEMPO model (Ghitza, 2011), the general architecture proposed by Giraud and Poeppel (2012), or a number of computational models (Hyafil, Fontolan, Kabdebon, Gutkin, & Giraud, 2015; Hovsepyan, Olasagasti, & Giraud, 2020; ten Oever & Martin, 2021). In such models, it is assumed that information channels at different timescales support and implement the temporal segmentation of speech signals. The main assumption concerns the theta range (4–8 Hz) which appears well suited for syllabic segmentation. The gamma range (25–50 Hz) is usually assumed to correspond to a temporally finer-grained classification of the signal input into phonetic events (phones, that may or may not

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.* ©2022 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

correspond to phonemic linguistic units per se). Such models feature sequences of decoders, at one or several linguistic levels, that are fed with segments of the speech signal. These models precisely describe processes in charge of opening and closing, in sequence, these decoders.

To the best of our knowledge, no model of how segmentation cues can be used to parse a sensory stream into decoders, sequentially, has been proposed in the probabilistic framework. In this paper, we propose original mathematical tools to address this issue. Based on coherence variables (Gilet, Diard, & Bessière, 2011; Bessière, Mazer, Ahuactzin, & Mekhnacha, 2013) and controlled coherence variables (Ginestet, Phénix, Diard, & Valdois, 2019), we define probabilistic constructs (portions of models that we call "Bayesian gates"), that use segmentation cues so as to parse a temporally continuous stream of sensory input into a sequence of "decoders" (perceptual accumulators of sensory evidence), so that each decoder only receives and analyzes a portion of the sensory input.

For generality purposes, we first mathematically define Bayesian gates in a small, abstract model (i.e., agnostic to its application domain), and show how Bayesian inference yields temporal segmentation of the sensory input. We then show application of Bayesian gates in a larger-scale model of speech perception (Nabé, Schwartz, & Diard, 2021), in which a temporal submodel provides temporal cues concerning syllable onsets, to be used by Bayesian gates to perform segmentation of the acoustic signal, and feed it sequentially into syllable decoders, to perform word recognition.

The rest of this paper is structured as follows. The next section first introduces the mathematical definition of Bayesian gates, in a simple and generic two-decoder case, and second, illustrates how they are applied in a larger-scale model of speech perception. Then, we present simulations and simulation results to illustrate how Bayesian gates perform the temporal segmentation of acoustic signal into a sequences of phone and syllable decoders, during speech perception.

Model

Controlled coherence variables for temporal gates

In the most simple case, we consider two decoders (noted 1 and 2 in subscripts in the following mathematical notations), each involving a representation of the sensory input (variables S) connected to an accumulator of perceptual evidence (variables P). Each decoder is connected to a mechanism built upon coherence variables (variables λ, one per decoder) and a control variable (G). This mechanism constitutes the "Bayesian gate" *per se*. This is itself informed by another submodel (variable C), in charge of controlling the temporal segmentation.

Given this architecture, the overall mechanism underlying Bayesian gates can be described as follows. The probability distribution over variable C acts as an event detector of temporal events from the sensory input (or predicting them from some other source of information). This probability distribution then acts as a signal to trigger Bayesian gates: the probability distribution over variable G represents a reference to the active Bayesian gate; when an event is detected, the current Bayesian gate is closed and the next one opens. This is performed mathematically by controlling coherence variables λ, which modulate the information flow from sensory input to the decoders, as if the architecture of the model were changed on the fly. We now describe how this mechanism is implemented mathematically.

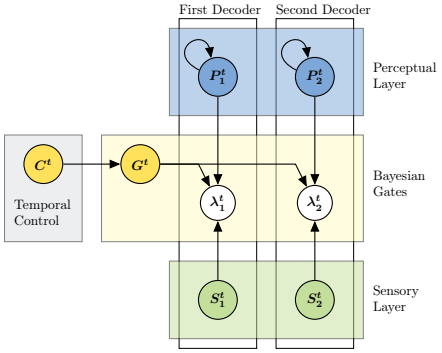


Figure 1: Graphical representation of the dependency structure of the probabilistic model for Bayesian gates. Nodes correspond to variables in the models, and edges represent dependencies, following the convention of notation of Bayesian Networks (self-looping arrows represent temporal dependencies between variable at time t and the preceding time step $t-1$). Colored boxes and text identify portions of the model (see text for details).

We consider a fine-grained time step (e.g., one time step for one millisecond), and note in superscripts time indices for all variables. Let us assume a time range from time instants 0 to T, and use the shorthand $X^{t_1:t_2}$ to denote the set of all variables X^t, with $t \in [\![t_1;t_2]\!]$. Therefore, the joint probability distribution to define the model is:

$$P(S_{1:2}^{1:T}\ P_{1:2}^{0:T}\ \lambda_{1:2}^{1:T}\ G^{1:T}\ C^{1:T})\ . \tag{1}$$

To define the joint probability distribution of Eq. (1), we use the dependency structure that is illustrated Figure 1. This decomposes the joint probability distribution into:

$$P(S_{1:2}^{1:T}\ P_{1:2}^{0:T}\ \lambda_{1:2}^{1:T}\ G^{1:T}\ C^{1:T}) = \tag{2}$$
$$P(P_{1:2}^{0}) \prod_{t=1}^{T} \left[\begin{array}{l} P(C^t) P(G^t \mid C^t) \\ \prod_{i=1}^{2} \left[P(S_i^t) P(P_i^t \mid P_i^{t-1}) P(\lambda_i^t \mid S_i^t\ P_i^t\ G^t) \right] \end{array} \right] .$$

Variables S_i^t and P_i^t, whatever i and t, have the same, discrete and finite arbitrary domain $\mathcal{D}$, which is the representation space of the perceptual dimension of interest. The probability distribution $P(S_i^t)$ is the "sensory model" that feeds into the Markov chain over variables $P_i^{1:T}$, defined by the temporal model $P(P_i^t \mid P_i^{t-1})$ and prior distribution $P(P_i^0)$. Although this is not necessary for defining the Bayesian gate mechanism, in the following, the terms $P(P_i^t \mid P_i^{t-1})$ are defined to

only feature information leak, without any structure (contrary to classical HMM based models, for instance). To do so, we define:

$$P([P_i^t = p^t] \mid [P_i^{t-1} = p^{t-1}]) = \begin{cases} \frac{1+leak}{1+|\mathcal{D}|leak} & \text{if } p^t = p^{t-1} \\ \frac{leak}{1+|\mathcal{D}|leak} & \text{otherwise,} \end{cases} \tag{3}$$

with $|\mathcal{D}|$ the cardinal of domain $\mathcal{D}$ and *leak* a parameter controlling information decay speed.

With such Markov chains over perceptual variables P, and sensory models over sensory variables S, connecting them directly would yield straightforward decoders: when sensory distributions $P(S_i^t)$ are informed (that is, they are different from uniform distributions), they are fed at each time step into the Markov chains, which thus operate as accumulators of perceptual evidence, and the probability distribution over variable P^t gradually peaks, as a function of t, on the most likely sensory hypothesis. On the contrary, when the sensory distributions $P(S_i^t)$ are uniform, no perceptual evidence is available (simulating the absence of stimulation), and the Markov chains gradually decay back to uniform distributions.

We consider that sensory streams, $S_1^{1:T}$ and $S_2^{1:T}$, are duplicates of the sensory input. In that case, the two decoders would be fed with the same sensory information, and thus, the probability distributions over variables $P_1^{1:T}$ and $P_2^{1:T}$ would also be identical. The purpose of Bayesian gates is exactly to parse out the sensory streams, so that one portion is fed into $P_1^{1:T}$, and another into $P_2^{1:T}$, so that perceptual decoders process different segments of the sensory streams.

To define Bayesian gates, we assume first that variable G^t, for any t, has a discrete and finite domain that maps to the number of decoders in the model. In our simple case, since we consider two decoders, the domain of G^t would be $\{1,2\}$. When $G^t = i$, this means that the i-th decoder is "open" (it receives sensory information), and all others are "closed" (they receive uniform distributions as input). To pilot the transfer of information between sensory and perceptual variables, we define the λ variables as controlled coherence variables (Nabé et al., 2021), that is to say, they are binary variables and:

$$\begin{aligned} &P([\lambda_i^t = 1] \mid [S_i^t = s^t]\,[P_i^t = p^t]\,[G^t = g^t]) \\ &= \begin{cases} 1 & \text{if } s^t = p^t \text{ and } g^t = i \\ 0 & \text{if } s^t \neq p^t \text{ and } g^t = i \\ 1/|\mathcal{D}| & \text{if } g^t \neq i\,. \end{cases} \end{aligned} \tag{4}$$

Finally, the Bayesian gate is connected to an external submodel, that provides cues about segmentation events. In the general case, it can be arbitrarily complex, and rely on any source of available information. Here, for simplicity, we represent this temporal control submodel with a single probability distribution, $P(C^t)$. We assume that C^t, for any t, is a Boolean variable, and that $P([C^t = True])$ represent the probability that a segmentation event is detected. If such an event is detected, the currently opened decoder should be closed, and the next one should be opened: this is implemented with probability distribution $P(G^t \mid C^t)$, defined by a Dirac distribution over G^t. Recall that variable G^t indexes decoders, so that $P(G^t \mid C^t)$ assigns probability 1 to "the next decoder" (i.e. decoder $i+1$, with i an internal parameter that tracks the currently opened gate).

All terms featured in Eq. (2) are described, so that the model is fully defined. We now show how the model provides the desired behavior for Bayesian gates. To do so, we consider computing the probability distribution over perceptual variable P_i^t, given event detection $C^{1:t}$ and sensory input stream $S_{1:2}^{1:t}$ (and assuming, for technical reasons, that λ variables are 1). We note this term:

$$Q_i^t = P(P_i^t \mid C^{1:t}\, S_{1:2}^{1:t}\, [\lambda_{1:2}^{1:t} = 1])\,. \tag{5}$$

To compute Q_i^t, applying Bayesian inference in the model yields:

$$\begin{aligned} Q_i^t &= P(P_i^t \mid C^{1:t}\, S_{1:2}^{1:t}\, [\lambda_{1:2}^{1:t} = 1]) \\ &\propto \left[P([G^t = i] \mid C^t)P(S_i^t) + P([G^t \neq i] \mid C^t)/|\mathcal{D}|\right] \\ &\quad \times \sum_{P_i^{t-1}} \left[P(P_i^t \mid P_i^{t-1})P(P_i^{t-1} \mid C^{1:t-1}\, S_{1:2}^{1:t-1}\, [\lambda_{1:2}^{1:t-1} = 1])\right]\,. \end{aligned}$$

To make this result readable, first, we note $\alpha_i = P([G^t = i] \mid C^t)$, so that $P([G^t \neq i] \mid C^t) = 1 - \alpha_i$, second, we recognize that the constant value $1/|\mathcal{D}|$ can be interpreted as the probability value of the uniform distribution over domain $\mathcal{D}$, noted $U_{\mathcal{D}}$, and third, we recognize that the last factor under the sum is the recurrence term, that is to say, the same computation as Q_i^t but at the previous time step, so that it can be noted as Q_i^{t-1}. We obtain:

$$Q_i^t \propto \left[\alpha_i P(S_i^t) + (1-\alpha_i)U_{\mathcal{D}}\right] \sum_{P_i^{t-1}} \left[P(P_i^t \mid P_i^{t-1})Q_i^{t-1}\right]\,. \tag{6}$$

We recognize here the usual form for Bayesian filtering in the large class of temporal probabilistic models (Russell & Norvig, 1995; Murphy, 2002, 2012): inside the sum, the recursive term is multiplied by the temporal model; the sum itself is then multiplied by a term, which takes here a particular form that merits attention. Indeed, $\alpha_i P(S_i^t) + (1-\alpha_i)U_{\mathcal{D}}$ is a weighted sum between the sensory model proper $P(S_i^t)$, and a uniform distribution. Weights are $\alpha_i = P([G^t = i] \mid C^t)$, that is to say, the probability that decoder i is open at time t, and $1 - \alpha_i = P([G^t \neq i] \mid C^t)$, that is, the probability that it is closed. In other words, when decoder i is open, the Bayesian gate between the temporal model and sensory distribution is open, so that $\alpha_i = 1$, and the temporal model "sees as sensory input" the distribution $P(S_i^t)$. On the other hand, when decoder i is closed, $\alpha_i = 0$, so that the temporal model "sees as sensory input" a uniform distribution. Multiplying by a uniform distribution has no effect (the uniform distribution is the neutral element of multiplication in the probabilistic setting), so that $\alpha_i = 0$ is equivalent to having no sensory input to process. This demonstrates that Bayesian gates, thanks to Eq. (6), act as desired and parse the sensory stream $P(S_i^t)$ so as to feed it only into the opened gate at time t.

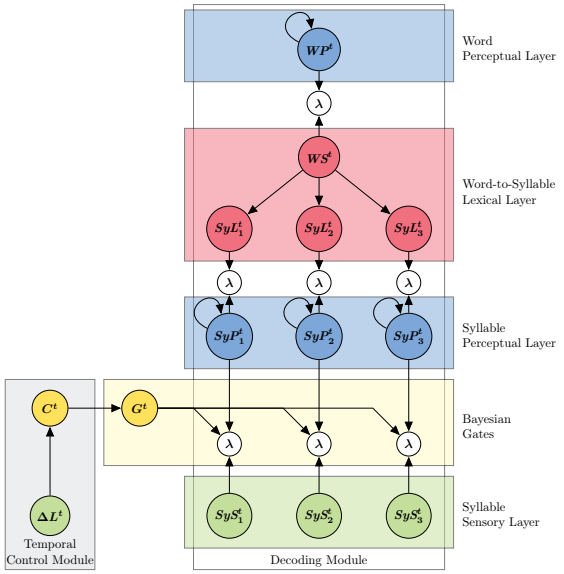


Figure 2: Graphical representation of the dependency structure of a simplified version of the COSMO-Onset model. Same graphical convention as in Figure 1.

Application to the COSMO-Onset model

Bayesian gates have been applied in a model of speech perception, COSMO-Onset (Nabé et al., 2021), that we briefly describe. The overall architecture of a simplified version of the COSMO-Onset model is graphically represented Figure 2. It is a hierarchical probabilistic model, and its overall dependency architecture is organized around two main modules: the decoding module, and the temporal control module.

The architecture of the decoding module is inspired by the classical interactive-activation models, such as the TRACE model (McClelland & Elman, 1986), and it is similar to those of recent models (Hovsepyan et al., 2020; Yildiz, von Kriegstein, & Kiebel, 2013). The decoding module is organized hierarchically with alternating layers of perceptual accumulation and lexical knowledge, from a pre-processing stage consisting of acoustic feature extraction (phone sensory layer), to syllable-to-phone knowledge and word-to-syllable knowledge, through phone, syllable and word perceptual models. Concerning temporal events and segmentation, the decoding module assumes that a word is composed of a sequence of syllables (in the experiments below, at most 3), each composed of a sequence of phones (at most 4). For simplicity, portions of the model involving phones are not represented in Figure 2, so that the input of the decoding module, in this simplified version, is considered at the syllabic level (variables $SyS^t_{1:12}$). In other words, in the following, we simplify the presentation of the model to consider that the sensory input would already inform about syllable identity (whereas, in the complete model, this involves an intermediary step to infer syllable from phone sequences, with phones inferred by an analysis of the acoustic content of the input signal).

The second module is the temporal control module, as in the simple model presented above. In this paper, we consider that the temporal control module builds a probability distribution over variable C^t, using increases of the acoustic intensity in the speech signal (variable ΔL^t, with L for "loudness"), as likely candidate events for syllabic onsets, and thus, salient time steps for speech segmentation. (Nabé et al. (2021) considered lexical knowledge about syllable duration, as top-down cues to complement the bottom-up cues based on syllable onset detection from the signal; we only consider bottom-up cues here.) This is the portion of the COSMO-Onset model that contains the Bayesian gates that we experimentally study below. However, we note that, in the full COSMO-Onset model, once a Bayesian gate detects a syllabic onset, not only does it close the current syllable decoder and open the next one, it also triggers a sequence of four phone decoders. These are opened and closed on a fixed schedule (every 50 time steps); however, it is entirely possible that this sequence is ended prematurely, whenever the next syllabic onset is detected.

Experiments

Materials

In the experiments we present here, the model is configured with a lexicon of known words comprised of 28 words, with 7 monosyllabic words, 14 bisyllabic words and 7 trisyllabic words. Syllables are Consonant-Vowel (CV) syllables composed of a plosive consonant followed by a vowel; the first syllable of a word can be a single Vowel (V) syllable. The considered consonants are /p/ and /t/, and the vowels are /a/, /i/ and /u/. Some examples of words in the lexicon are "a", "pa", "pi", "apa", "patu", "apata", "iputu", "tapatu"; the complete list is provided elsewhere (Nabé et al., 2021, Table 2).

For every word in the lexicon, a definition of their composition at the acoustic and phonetic level is given. These correspond to synthetic, "toy-like" realization, in terms of duration, spectral content and loudness profiles. Concerning duration, every phone (consonant or vowel) lasts 50 time steps, to correspond to 50 ms of physical time. In CV syllables, a transitional phone (noted "@") is inserted between the consonant and vowel, that also lasts 50 time steps. V syllables at the beginning of words last 100 time steps. Therefore, syllable boundaries do not always occur at the same time step, or on multiples of the duration of a CV syllable (150); furthermore, the model is not given information about phone or syllable duration, and detects syllabic onset events from the signal, whenever they might occur. Therefore, setting 50 time steps as the base duration for phones is merely a convenience, to keep the toy lexicon simple, and not a limit of the model.

The spectral contents of vowels are defined by their characterization in the (F1, F2) plane, around usual prototypical values (see Figure 3, left panel). Since we only consider the /p/ and /t/ stop consonants, they are also defined in the (F1, F2) plane. Each phone consists in a repetition of a (F1, F2) point

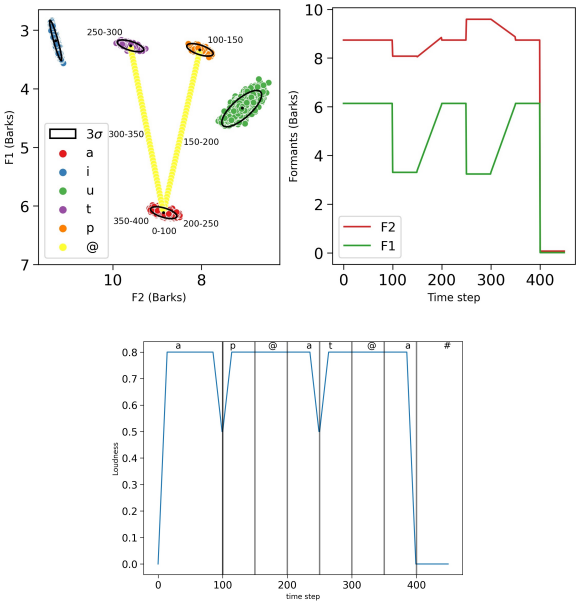


Figure 3: Spectral contents (top) and loudness profiles (bottom) of synthetic stimuli used in the experiment. Top, left: regions of the (F1, F2) plane (x- and y-axes, in Barks) for each phone considered. Ellipses and clouds of points indicate the possible variations of realizations for each phone. The indicated time steps correspond to the complete trajectory for word "apata": in the "a" region between time step 0 and 100, in the "p" region between time step 100 and 150, and so on. Top, right: trajectories of F1 and F2 values (y-axis) as a function of time, for word "apata". Bottom: Loudness values (y-axis) as a function of time, for word "apata".

for 50 iterations, with this point drawn in an ellipse around prototypical values. Transition phones "@" are linear trajectories in the formant space, between the preceding consonant and following vowel.

Finally, the synthetic stimuli are also defined by loudness profiles, to mimic the envelope of the speech signal. A scalar, positive value is defined at each time step. It is constant (0.8) "inside" phones, except for an energy rise (from 0 to 0.8) at the beginning of words, an energy decrease (0.8 to 0) at the end of words, and an energy dip at syllabic boundaries (from 0.8 to 0.5 towards the end of syllable i, and from 0.5 to 0.8 at the beginning of syllable $i+1$). Synthetic stimuli end with 50 time steps of "simulated silence", noted "#", with loudness set to 0 (and formants set to (0,0) also, as a convention).

Figure 3 illustrates the spectral content and loudness profile for the synthetic stimulus corresponding to word "apata".

Methods

We have conducted an experiment to study the effect of Bayesian gates and temporal segmentation during word recognition. More precisely, we assessed the robustness of the model to temporal misalignment, by performing a simulation experiment in which we manually inserted a delay between onset detection and its use for opening and closing Bayesian gates. In other words, the model would compute onset detection in a normal fashion (term $P(C^t \mid \Delta L^t)$), but its output would be temporally delayed before being transferred to variable G^t (the term $P(G^t \mid C^t)$ becomes $P(G^{t+delay} \mid C^t)$).

We have performed word recognition on all words of the lexicon, and varied the delay between -75 to +75 time steps (steps of 5 iterations). For all words and all delays, we have measured the probability assigned by the model to the input word (i.e., correct recognition probability) at the final iteration. The condition where the delay is 0 provides a base-case performance for the model.

Results

To illustrate how Bayesian gates segment an acoustic signal into a sequence of decoders in the COSMO-Onset model, we first describe the behavior of the model on a typical example. From the word recognition experiment, we consider the stimulus signal corresponding to word "apata" (see Figure 3). Figure 4 shows the evolution of word and syllable probabilities, in each of the three syllable decoders, in the base-case condition (delay is 0).

We observe that the three syllable decoders are activated sequentially, that each is fed portions of the acoustic input, leading to correct syllable recognition. We also observe that the third decoder, initially (around iterations 250 to 260), increases probability of syllable "a", erroneously. This is due to a slight misalignment: onset detection triggered slightly early, leading the third decoder to process a portion of the acoustic signal of the end of the second syllable (the end portion of the "pa" of "apata"). Later on, this is corrected, as the input enters the "t" of "apata", and the third decoder correctly assigns high probability that the third syllable would be "ta" (red curve, bottom plot of Figure 4). We also observe that the probability distribution over words, as time increases, narrows down competing hypotheses, and also yields, at the end, very high probability for the correct word "apata". Overall, this simulation illustrates that the model and the Bayesian gates mechanism behave as expected and yield correct syllable segmentation and recognition, and thus, correct word recognition on input "apata". This illustrative example also suggests that the model would be robust to a small misalignment between temporal events in the acoustic signal and the opening and closing of syllable decoders.

Experimental results for the whole experiment, in which the manually-inserted delay is varied systematically, are shown Figure 5. We observe an inverted-U shaped plot, with the probability for the correct word maximal when the delay is 0 or +5 iterations (probabilities differ at the third decimal), and very close to maximal when the delay is +10 iterations. For other delay values, we observe that performance sharply decreases. We have analyzed results independently for monosyllabic, bisyllabic and trisyllabic words (not shown). Monosyllabic words are overall better recognized, and performance

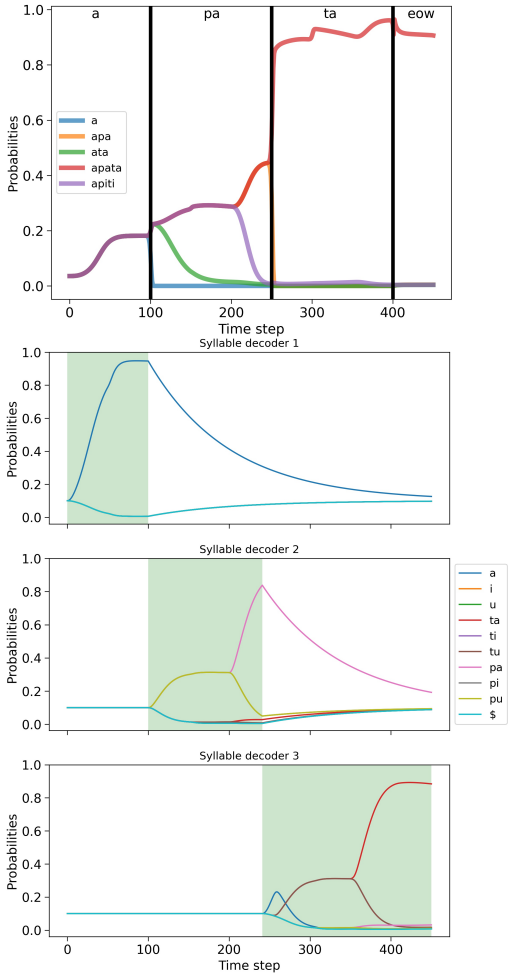


Figure 4: Simulation of word recognition in COSMO-Onset on input "apata". Top plot: evolution of probabilities of the most likely word hypotheses (y-axis) as a function of simulated time (x-axis). The vertical black lines indicate time steps when syllable onsets were detected. Top annotations recall the contents of acoustic input. Bottom three plots: evolution of probabilities of syllables (y-axis) as a function of simulated time (x-axis), in the three syllable decoders. Colored intervals show when the Bayesian gate of each decoder was open. (Some curves are partly superposed.)

is more robust; this, of course, is due to the fact that monosyllabic word recognition is only dependent on a single syllabic onset detection. Result patterns for bisyllabic and trisyllabic words are very similar to the global results of Figure 5.

Overall, our experimental results suggest that, when the COSMO-Onset model processes our synthetic stimuli, there is a small temporal tolerance, for which performance is preserved. However, performance is worse for large delays, which confirms that a proper alignment of syllabic decoders with the acoustic signal is central for word recognition.

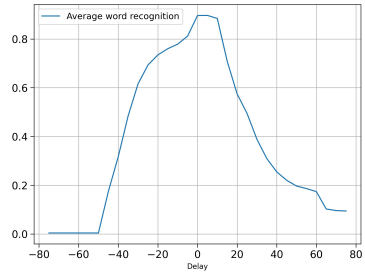


Figure 5: Average probability for the correct word (y-axis) in COSMO-Onset simulating word recognition, over all words of the lexicon, as a function of a manually-imposed delay between onset detection and their use for opening and closing Bayesian gates (x-axis, in iterations).

Discussion

We have proposed a novel probabilistic construct, called "Bayesian gates", to segment a sensory stream so that each decoder in a sequence of perceptual decoders is fed with a portion of the sensory stream. We have defined Bayesian gates and explored their mathematical properties in a simple model. Then, we have shown how Bayesian gates have been applied in a speech perception model. In this model, syllable onset detection is used as the signal controlling Bayesian gates, to feed a sequence of syllable decoders. These perform syllable recognition, upon which word recognition relies. On synthetic stimuli, we have experimentally shown that Bayesian gates fulfill their role of temporal segmentation, and that they are robust to slight temporal misalignment between syllable onsets and the activation of syllable decoders.

In the current paper, we have defined $P(G^t \mid C^t)$ to be a Dirac distribution, indexing in an all-or-nothing fashion the syllabic decoder to be opened. This implies that, when word recognition processing unfolds, at each time step, the states of syllable decoders are known with certainty. In other words, a single decoding trajectory is computing, with decoders either "fully opened" or "fully closed". Relaxing the Dirac assumption would allow representing probability distributions with uncertainty instead. In that case, gates would simultaneously be opened and closed, in proportions quantified by probabilities. In principle, this should allow, as in HMM-based decoding, to compute simultaneously several decoding trajectories, and, possibly, to correct past errors with future information (decreasing probabilities of trajectories when they lead to less likely decoding paths). Assessing the computational cost induced by such a mechanism, its possible performance gain, and its cognitive plausibility, is part of ongoing research.

Overall, we have described a segmentation mechanism, suitable for syllabic parsing, based on sensory cues extracted in a bottom-up manner from the acoustic stimulus. Current work aims at exploring whether theta oscillations could implement our mechanistic model, and whether other information channels, maybe in the beta band, could implement complementary, top-down, lexically driven prediction.

Acknowledgments

This work is supported by the French National Research Agency in the framework of the Investissements d'avenir program (ANR-15-IDEX-02; Ph.D. grant to MN from Université Grenoble Alpes ISP project Bio-Bayes Predictions). Authors also acknowledge additional support by the Auvergne-Rhône-Alpes (AURA) Region (PAI-19-008112-01 grant). This work has also been partially supported by the Multidisciplinary Institute of AI (MIAI) @ Grenoble Alpes (ANR-19-P3 IA-0003).

References

Bessière, P., Mazer, E., Ahuactzin, J. M., & Mekhnacha, K. (2013). *Bayesian programming*. Boca Raton, Florida: CRC Press.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology*, *19*(2), 141–177.

Gales, M., & Young, S. (2008). The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing*, *1*(3), 195–304.

Ghitza, O. (2011). Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, *2*(130), 1–13.

Gilet, E., Diard, J., & Bessière, P. (2011). Bayesian action-perception computational model: Interaction of production and recognition of cursive letters. *PLoS ONE*, *6*(6), e20387.

Ginestet, E., Phénix, T., Diard, J., & Valdois, S. (2019). Modeling the length effect for words in lexical decision: The role of visual attention. *Vision Research*, *159*, 10–20.

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517.

Girin, L., Leglaive, S., Bie, X., Diard, J., Hueber, T., & Alameda-Pineda, X. (2021). Dynamical variational autoencoders: A comprehensive review. *Foundations and Trends in Machine Learning*, *15*(1–2), 1–175.

Hovsepyan, S., Olasagasti, I., & Giraud, A.-L. (2020). Combining predictive coding and neural oscillations enables online syllable recognition in natural speech. *Nature Communications*, *11*, 3117.

Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., & Giraud, A.-L. (2015). Speech encoding by coupled cortical theta and gamma oscillations. *eLife*, *4*, e06213.

Lotto, A. J., & Holt, L. (2000). The illusion of the phoneme. In S. J. Billings (Ed.), *The panels (Chicago linguistic society)* (pp. 191–204). Carnegie Mellon University.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

Murphy, K. (2002). *Dynamic Bayesian networks: Representation, inference and learning*. Ph. D. thesis, University of California, Berkeley, Berkeley, CA.

Murphy, K. (2012). *Machine learning: A probabilistic perspective*. Cambridge, MA: MIT Press.

Nabé, M., Schwartz, J.-L., & Diard, J. (2021). COSMO-Onset: a neurally-inspired computation model of spoken word recognition, combining top-down prediction and bottom-up detection of syllabic onsets. *Frontiers in Systems Neuroscience*, *15*, 653975.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE Trans. on ASSP*, *77*(2), 257–285.

Russell, S., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Englewood Cliffs, New Jersey: Prentice Hall Series in Artificial Intelligence.

Schlittenlacher, J., & Ellermeier, W. (2021). Continuous magnitude production of loudness. *Frontiers in Psychology*, *12*, 635557.

Stevens, S. S. (1955). The measurement of loudness. *The Journal of the Acoustical Society of America*, *27*(5), 815–829.

ten Oever, S., & Martin, A. E. (2021). An oscillating computational model can track pseudo-rhythmic speech by using linguistic predictions. *eLife*, *10*, e68066.

Yildiz, I. B., von Kriegstein, K., & Kiebel, S. J. (2013). From birdsong to human speech recognition: Bayesian inference on a hierarchy of nonlinear dynamical systems. *PLoS computational biology*, *9*(9), e1003219.

Zwicker, E., & Sharf, B. (1965). A model of loudness summation. *Psychological Review*, *72*(1), 3–26.