COSMO-Onset : a Bayesian, neurally inspired model of speech perception combining bottom-up envelope processing and top-down predictions for syllabic segmentation

Mamady Nabe

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Informatique
Unité de recherche : Laboratoire de Psychologie et Neuro Cognition

COSMO-Onset : un modèle Bayésien de perception de la parole, neuro-inspiré, combinant un traitement bottom-up de l'enveloppe du signal et des prédictions temporelles top-down pour la segmentation syllabique

COSMO-Onset: a Bayesian, neurally inspired model of speech perception combining bottom-up envelope processing and top-down predictions for syllabic segmentation

Présentée par :

Mamady NABE

Direction de thèse :

Julien DIARD — Directeur de thèse
Chargé de recherche HDR, CNRS Délégation Alpes

Jean-luc SCHWARTZ — Co-directeur de thèse
Directeur de recherche, CNRS Délégation Alpes

Rapporteurs :

NOEL NGUYEN
Professeur des Universités, AIX-MARSEILLE UNIVERSITE

FREDERIC BIMBOT
Directeur de recherche, CNRS BRETAGNE ET PAYS DE LA LOIRE

Thèse soutenue publiquement le **14 mars 2023**, devant le jury composé de :

JULIEN DIARD Chargé de recherche HDR, CNRS DELEGATION ALPES	Directeur de thèse
JEAN-LUC SCHWARTZ Directeur de recherche, CNRS DELEGATION ALPES	Co-directeur de thèse
NOEL NGUYEN Professeur des Universités, AIX-MARSEILLE UNIVERSITE	Rapporteur
FREDERIC BIMBOT Directeur de recherche, CNRS BRETAGNE ET PAYS DE LA LOIRE	Rapporteur
LAURENT GIRIN Professeur des Universités, GRENOBLE INP	Président
ITSASO OLASAGASTI Ingénieur docteur, Université de Genève	Examinatrice
OKKO RASANEN Professeur associé, Tampereen Yliopisto	Examinateur

Summary

Neurocognitive speech perceptual processing is classically conceived as a hierarchy of computations – typically including acoustic or multi-sensory feature extraction, pre-lexical categorization, lexical access, prosodic and syntactic integration, up to final comprehension stages. It is increasingly considered that neural communication within and across these various stages is based on synchronization processes and operates thanks to chunking and selection mechanisms exploiting neural oscillatory dynamics at various frequencies.

In contrast to classical models of speech perception such as the TRACE or SHORTLIST models, which achieve segmentation solely through the decoding of the spectro-temporal content of the speech input, recent neuroscience research in speech perception advocates for a clear separation between two processing pathways: a decoding pathway and a temporal control pathway. The latter proposal has given rise to several neuro-computational models, which, for segmentation, rely solely on the processing of the acoustic envelope enabling syllabic rhythm tracking from the speech signal. In this sense, they are entirely "bottom-up" segmentation models.

However, several studies have shown that reliable speech perception can not be achieved only through bottom-up processes. For instance, clear evidence for the role of top-down temporal predictions has been provided by Aubanel and Schwartz (2020). Their study showed that speech sequences embedded in noise were better processed and understood by listeners when they were presented in their natural, irregular timing than in timing made isochronous, without changing their spectro-temporal content. The strong benefit in intelligibility displayed by natural syllabic timing, both in English and in French, was interpreted by the authors as evidence for the role of top-down temporal predictions for syllabic parsing.

The objective of the present thesis is to address the question of the fusion of bottom-up and top-down processes for speech syllabic segmentation. Our contribution is the COSMO-Onset model, a Bayesian hierarchical model of speech perception, involving a speech segmentation module with an original top-down mechanism for syllabic onset prediction, involving lexical temporal knowledge. We use the model to explore the respective roles of bottom-up envelope processing and top-down linguistic predictions and how they can be efficiently combined for syllabic segmentation. On a first set of experiments on simplified, synthetic stimuli, we show that while purely bottom-up onset detection is sufficient for word recognition in nominal conditions, top-down prediction of syllabic onset events allows overcoming challenging adverse conditions, such as when the acoustic envelope is degraded, leading either to spurious or missing onset events in the sensory signal. On a second set of experiments on real speech stimuli from the Aubanel and Schwartz (2020) experiment, we show that the COSMO-Onset model succesfully accounts for the complementary roles of isochrony and naturalness in speech perception in noise.

Résumé

Le traitement neurocognitif de la perception de la parole est classiquement conçu comme une hiérarchie de calculs - comprenant typiquement l'extraction de caractéristiques acoustiques ou multi-sensorielles, la catégorisation pré-lexicale, l'accès lexical, l'intégration prosodique et syntaxique, jusqu'aux étapes finales de compréhension. On considère de plus en plus que la communication neuronale au sein et entre ces différentes étapes est basée sur des processus de synchronisation et fonctionne grâce à des mécanismes de découpage et de sélection exploitant la dynamique oscillatoire neuronale à diverses fréquences.

Contrairement aux modèles classiques de perception de la parole tels que les modèles TRACE ou SHORTLIST, qui réalisent la segmentation uniquement par le décodage du contenu spectro-temporel de l'entrée de la parole, les recherches récentes en neurosciences sur la perception de la parole préconisent une séparation claire entre deux voies de traitement : une voie de décodage et une voie de contrôle temporel. Cette dernière proposition a donné lieu à plusieurs modèles neuro-computationnels qui, pour la segmentation, reposent uniquement sur le traitement de l'enveloppe acoustique permettant le suivi du rythme syllabique à partir du signal de parole. En ce sens, il s'agit de modèles de segmentation entièrement "bottom-up".

Cependant, plusieurs études ont montré qu'une perception fiable de la parole ne peut être obtenue uniquement par des processus "bottom-up". Par exemple, des preuves claires du rôle des prédictions temporelles "top-down" ont été fournies par Aubanel and Schwartz (2020). Leur étude a montré que les séquences vocales intégrées dans le bruit étaient mieux traitées et comprises par les auditeurs lorsqu'elles étaient présentées dans leur timing naturel et irrégulier que dans un timing rendu isochrone, sans changer leur contenu spectro-temporel. Le fort bénéfice en intelligibilité affiché par le timing syllabique naturel, tant en anglais qu'en français, a été interprété par les auteurs comme une preuve du rôle des prédictions temporelles descendantes pour l'analyse syllabique.

L'objectif de la présente thèse est d'aborder la question de la fusion des processus "bottom-up" et "top-down" pour la segmentation syllabique de la parole. Notre contribution est le modèle COSMO-Onset, un modèle hiérarchique bayésien de la perception de la parole, impliquant un module de segmentation de la parole avec un mécanisme descendant original pour la prédiction de l'apparition syllabique, impliquant des connaissances temporelles lexicales. Nous utilisons le modèle pour explorer les rôles respectifs du traitement "bottom-up" de l'enveloppe et des prédictions linguistiques "top-down", et comment ils peuvent être combinés efficacement pour la segmentation syllabique. Dans une première série d'expériences sur des stimuli synthétiques simplifiés, nous montrons que si la détection purement "bottom-up" du début de la parole est suffisante pour la reconnaissance des mots dans des conditions nominales, la prédiction "top-down" des événements syllabiques du début de la parole permet de surmonter des conditions défavorables difficiles, comme lorsque l'enveloppe acoustique est dégradée, ce qui conduit à des événements de début de parole parasites ou manquants dans le signal sensoriel. Sur une deuxième série d'expériences sur des stimuli de parole réels provenant de l'expérience d'Aubanel and Schwartz (2020), nous montrons que le modèle COSMO-Onset rend compte avec succès des rôles complémentaires de l'isochronie et du naturel dans la perception de la parole dans le bruit.

Remerciements – Acknowledgements

Voici venu le moment redoutable. J'ai longtemps hésité en réalité à me décider pour écrire ces quelques lignes de remerciement. D'un côté, j'avais peur de ne pas pouvoir suffisamment "doser" le contenu, tant il y a de gens que j'aimerais remercier, et davantage pour certain.e.s particulièrement. Mais de l'autre côté, c'est quand même une bien unique occasion d'exprimer enfin ces remerciements en clair pour en garder une trace somewhere.

First, I would like to sincerely thank Frédéric Bimbot and Noël Nguyen who accepted to be reviewers and evaluate this thesis. I thank the examiners Okko Räsänen, Itsaso Olasagasti, and Laurent Girin for being members of the jury. I thank them all for taking the time to be interested in my work, and for the very interesting and challenging discussions during my defense.

Je continue pour le reste en Français. Je remercie très spécialement mes deux directeurs de thèse pour l'opportunité et la confiance qu'ils m'ont accordé. Je me rappelle encore de ce premier jour de notre contact en personnes. C'était pour un entretien sur mes motivations pour la thèse. Comme le diraient certains, je m'étais mis sur mon 31, avec un blaser assorti d'un soulier qui paraissait chic mais était en réalité du low-cost. À mon arrivé dans le hall du bâtiment du laboratoire de psychologie et de neurocognition (LPNC), je vois venir un monsieur en short et chemise, avec des sandales aux pieds. Il s'agissait de Jean-Luc. Il faut dire qu'il était mieux habillé que moi, car nous étions dans une journée de chaleur, et par conséquent il devait se sentir nettement mieux dans son accoutrement que moi. Comme perdu dans un labyrinthe, un autre monsieur m'adressa la parole, en me demandant si j'étais bien Mamady. Il était habillé en jean et pull, avec des chaussures aux airs de marcheur chevronné. C'était Julien. Il m'a ensuite conduit dans une salle de réunion où devait se passer l'entretien. Quelques semaines plus tard, je savais que ce labyrinthe allait être mon lieu de travail de doctorant, et que les deux monsieurs susmentionnés, mes chers encadrants. Chaque jour passé à côté de ces deux était une bénédiction. Vous savez que vous êtes bien tombé lorsque le simple fait de mentionner aux autres collègues

que vous êtes encadré par Julien et Jean-Luc, vous recevez tout de suite une expression de joie et de gratitude. Je sais ces quelques lignes sembleront trop pour Julien, mais je me le permets tout de même. À vous deux, je vous remercie pour tout l'accompagnement sur les plans scientifique, professionnel et humain que vous n'avez cessé de m'accorder tout le long de ces années de thèse. Pour votre disponibilité, votre bienveillance, votre implication sans commune mesure, votre sympathie, et votre collaboration, je vous dis un grand merci. J'aurai pu continuer à vous amender ces phrases en exprimant ce que je trouve unique chez chacun de vous, par exemple sur le fait que Julien est très pointilleux lorsqu'il s'agit de rédaction pour des questions de formulation ou de mise en forme, ou encore pour sa qualité à facilement répérer les typos, de quelque nature que ce soit. Tu m'as appris à être plus incisif dans mes présentations. Même si il a fallu attendre ma soutenance pour que je m'en sépare de mes malheureux tics. Quant à Jean-Luc, j'ai toujours été impressionné par son "éthique" de travail. Quoiqu'il en soit, ce roc de la science ne finit jamais ses journées probablement pas avant 23h57 précise. À vous deux, vous m'avez appris à être plus rigoureux, et vous avez toujours encouragé le curieux en moi. Pour tout ça et le reste que je n'ai pas mentionné ici, soit par omission involontaire, soit par omission volontaire pour éviter d'allonger ces remerciements, je vous remercie du fond du coeur.

Je retourne à un ton plus "commode et professionnel", dans le souci d'éviter avoir des remerciements lourds. Pour les personnes qui suivent, vous comprendrez donc toute ma frustration, mais je sais que vous ne m'en voudrez pas. Je tiens à remercier tout le personnel permanent du LPNC, spécialement l'équipe administrative composée de Sanie, Thierry, Guylaine et Claire. Je n'oublie pas ceux du GIPSA-Lab, et spécialement ceux de l'équipe PCMD et de la chaire "Parole" de l'institut multidisciplinaire de l'intelligence artificielle (MIAI).

Je remercie tous les collègues non permanents (doctorants et post-doctorants) d'ici ou d'ailleurs pour toutes les interactions humaines et sociales au labo comme en dehors: Célise, Merrick, Méline, Ali, Rémi, Élie, Mathieu, Anna, Maryam, Lucie, Thomas, Marc-Antoine, Cynthia, François, Jérémie, Yoan, Aurore, Sam, Clément, Olivier, Adeline, Wilfried, Émilie, Juliette, et très certainement tous les autres que je n'ai pas cités. Je remercie en particulier tous les membres du bureau E115 durant ces années de thèse. Aux anciens qui ont fait un petit bout de chemin avec moi: Benjamin, Yannick, Mike. Et aux charmantes demoiselles qui ont été là jusqu'à la soutenance: Lise, Milèna, Alexandra, Stéphanie, Inès. Je vous remercie très sincèrement pour tous les fous rires et les moments inoubliables qui resteront à jamais gravés dans ma mémoire. Vos bonnes humeurs, sympathies, et discussions sérieuses tout comme "out of this world" m'auront bien accompagnées pendant les temps de cette thèse. Au lieu de me sentir comme l'intrus de la bande, et par conséquent créer ma bande à part, vous avez su égayé mes journées,

que ce soit dans les moments difficiles ou dans les moments pas si difficiles. Pour tout ça alors, one love for life.

Je remercie également tous les amis d'ici et d'ailleurs pour leur soutien inconditionnel durant toute ma thèse.

Pour finir, un merci spécial à ma famille, mes parents et mes frères et soeurs qui m'ont soutenu de près ou de loin. Cette thèse est aussi la vôtre.

Contents

List of Figures

List of Tables

Introduction

Speech perception and speech segmentation

Speech is achieved by letting the airflow from our lungs through our mouth and nasal cavity. This air stream is controlled by organs such as the tongue, lips, jaw, and larynx. It produces an acoustic wave, which conveys information about what is considered as a "first level of articulation/combination" (Martinet, 1960) that are phonemes embedded inside syllables. These elements aggregate into a higher second level of articulation which are the words that ensure contact with the external world (Berwick et al., 2013; Phillips, 2003; Smith, 2006). We then use our creativity to combine various words to form meaningful sentences, obeying specific syntactic constraints and embedded in adequate prosody (Jackendoff, 2003).

The hierarchical construct of speech, linearized as a continuous acoustic signal, must be processed, also hierarchically, by a perceptual system in order to understand the message it conveys (Benesty et al., 2008; Juang & Chen, 1998; O'shaughnessy, 2000; Rabiner, Schafer, et al., 2007). This ultimate goal involves a series of categorization processes that enable the association of continuous signals with discrete representations at various linguistic levels (Samuel, 2011; Werker & Tees, 1992). Crucially, such categorization processes require solving in some way the fundamental problem of segmentation, enabling to parse the continuous signal into discrete units, identifying relevant temporal events such as syllable boundaries (Miller & Eimas, 1995; Paget, 2013; Pisoni, 1985; Samuel, 2011).

Neural oscillation-based models of speech segmentation

Classical speech perception models inspired by interaction-activation processes such as TRACE or SHORTLIST (McClelland & Elman, 1986; McClelland & Rumelhart, 1981; Norris, 1994; Norris & McQueen, 2008), as well as automatic speech recognition models based on pattern-matching, such as Hidden Markov

Models (HMMs) or Deep Neural Networks (DNNs) (Gales & Young, 2008; Girin et al., 2021; Hinton et al., 2012; Nassif et al., 2019; Rabiner, 1989; Young et al., 2002), achieve segmentation through decoding the speech input directly from its spectro-temporal content. To do so, they rely on computational processes associating phonetic-prosodic, lexical and syntactic-semantic knowledge.

However, recent studies in speech neuroscience focusing on speech perception suggest a distinction between segmentation and decoding processes. Through synchronization processes between different populations of neurons operating in different frequency bands (typically the gamma band within 40–100 Hz, the theta band within 4–8 Hz, and the delta band within 1–3 Hz), the human brain would exploit neuronal oscillations to perform the temporal segmentation of incoming acoustic signals (Buzsaki, 2006; Buzsáki & Draguhn, 2004; Ding et al., 2016; Engel & Singer, 2001; Fries, 2015; Ward, 2003).

Although still a matter of debate, there is a growing consensus on the potential causal role of brain rhythms not only in the perception and understanding of speech (Poeppel & Assaneo, 2020) but also in language acquisition (Goswami, 2022). Two influential models relating speech perception and brain rhythms have been proposed by Ghitza (2011) and Giraud and Poeppel (2012), both suggesting similar mechanisms. In the latter for instance, the speech input would initially be parsed according to syllabic rhythm thanks to neural oscillatory processes in the theta band (4–8 Hz). Inside syllabic chunks, the acoustic spectro-temporal analysis would be conveyed by gamma oscillations at around 40 Hz. Further prosodic/syntactic parsing and binding would rely on lower frequency processes in the delta range (1–3 Hz). Of importance here, several studies have particularly shown the regularity of the syllabic rhythms over languages in the world (Cutler, 1994; Greenberg et al., 2003; Pellegrino et al., 2011; Ramus et al., 1999), which would be associated with the intrinsic neuronal properties of the theta band (Ding et al., 2017; Goswami & Leong, 2013).

These theoretical proposals gave rise to a number of recent neuro-computational models of speech perception exploring the possibilities offered by neural oscillations to address issues related to speech segmentation (Hovsepyan et al., 2020; Hyafil, Fontolan, et al., 2015; Räsänen et al., 2018; Yildiz et al., 2013). The common point between all these models is that they use a sensory, input-driven, and hence bottom-up approach, where slow modulations of the speech signal envelope would be tracked by endogenous cortical oscillations. This enables parsing speech into intermediate speech units such as syllables, which would be the pivot decoding unit within the continuous acoustic speech stream (Greenberg, 1998; Grosjean & Gee, 1987; Kolinsky et al., 1995; Meynadier, 2001; Rosen, 1992).

The role of top-down information in speech segmentation

In ideal conditions, where listeners would not face any speech degradation, these entirely bottom-up models are expected to perform adequately, relying on their oscillatory nature and adaptive ability to track the speech rhythm in a large range around natural speech rhythm.

However, in practice, they show a rather limited performance which could be due to their specific sensory-driven nature. Indeed, several studies have shown the importance of feedback processing from higher-order cortical regions (Bastos et al., 2015; Fontolan et al., 2014), which may be involved in higher-order speech processing stages such as syntactic and semantic levels or contextual information integration on a relatively longer temporal range (Pefkou et al., 2017). It seems likely that top-down predictions exploiting the listener's knowledge of the timing of natural speech (e.g., lexical or prosodic information) could improve the efficiency of purely bottom-up segmentation (M. H. Davis & Johnsrude, 2007; Kösem & Van Wassenhove, 2017; Meyer, 2018; Zekveld et al., 2006).

Recently, clear evidence for the role of top-down timing predictions has been provided by Aubanel and Schwartz (2020). Their study showed that speech sequences embedded in a large level of noise were better processed and understood by listeners when they were presented in their natural, irregular timing than in timing made isochronous without changing their spectro-temporal content. The strong benefit in intelligibility displayed by natural syllabic timing, both in English and in French, was interpreted by the authors as evidence for the role of top-down temporal predictions for syllabic parsing.

In the field of psycho-linguistics also, since the pioneer development of the TRACE model (McClelland & Elman, 1986), the question of the role of feedback processes in speech perception and comprehension has been the focus of intense discussions (McClelland et al., 2006; Norris et al., 2016), and led to many developments in the Bayesian framework (Hohwy, 2017; Kamper et al., 2017). Recent findings confirm that recurrence plays a crucial role in perceptual processing in the human brain (e.g., Donhauser & Baillet, 2020; Kietzmann et al., 2019; Spoerer et al., 2020).

All these theoretical claims are compatible with the predictive coding framework (Friston, 2005; Friston & Kiebel, 2009; Rao & Ballard, 1999), which hypothesizes that the brain is inherently predictive, exploiting internal states to make inferences about upcoming sensory data. This framework provides an interpretation of neuronal activity in which top-down predictions would be interleaved and integrated with bottom-up sensory processing. Top-down information from various stages of the speech perception process would be fed back

to lower processing stages, possibly exploiting the beta band (15–20 Hz) which is assumed to be a relevant channel for providing such descending predictions (Arnal, 2012; Arnal & Giraud, 2012; Cope et al., 2017; Engel & Fries, 2010; Rimmele et al., 2018; Sohoglu et al., 2012). However, the exact manner in which top-down, feedback processes interact with bottom-up, feedforward processes remains unclear.

The fusion of bottom-up information and top-down knowledge for speech perception

In fact, the question of the combination of bottom-up information extraction and top-down predictions from higher linguistic levels is actually not new. It is at the heart of all modern speech recognition architectures, such as the classical Hidden Markov Models (HMM) in which bottom-up acoustic cues are associated with top-down state transition probabilities in phonetic decoding or word recognition (Gales & Young, 2008; Rabiner et al., 1989) or more sophisticated architectures such as hierarchical HMMs (Murphy, 2002) or multi-scale HMMs enabling to incorporate hierarchical linguistic structures in language processing (Eyigöz et al., 2013). It is also central in recent neural speech recognition models, including recurrent architectures implementing top-down feedback in the decoding process (Graves et al., 2013); see a recent review by C. Kim et al. (2020).

Still, while the importance of top-down predictions has been largely discussed in the literature, it has been mainly focused on the mechanisms involved in the decoding process, and not on the segmentation process per se. And as far as models of speech perception based on neural oscillations are concerned, there is currently, to the best of our knowledge, no neuro-computational model incorporating top-down knowledge capable of accounting for behavioral data highlighting the role of top-down predictions in speech processing and understanding such as the ones provided in the study by Aubanel and Schwartz (2020).

Thesis objective

The objective of this thesis is to address the question of the fusion of bottom-up and top-down processes for speech syllabic segmentation. We approach this question in a Bayesian computational framework, which enables us to efficiently introduce, conceptualize and compare computational processes expressed in a unified probabilistic formalism (Bessière et al., 2008).

For this aim, we will explore the respective roles of bottom-up envelope processing and top-down linguistic predictions and how they can be efficiently combined for syllabic segmentation. A specific focus will be set on exploring

how top-down knowledge could help to overcome the impairments of bottom-up processing systems in the context of speech perception when the signal is degraded.

To address these questions, we introduce COSMO-Onset, a variant of the COSMO framework developed over the years to simulate speech communication processes in a perceptuo-motor framework (Barnaud, Bessière, et al., 2018; Laurent et al., 2017; Moulin-Frier et al., 2015; Moulin-Frier et al., 2012; Patri et al., 2015). The present variant does not incorporate at this stage the whole perceptuo-motor loop developed and studied in previous COSMO papers. Instead, it concentrates on the auditory pathway, detailing two mechanisms of interest for the present study: first, a hierarchical decoding process combining the phonetic, syllabic, and word levels, and, second and most importantly in the present context, a syllabic parsing mechanism based on event detection, operating on the speech envelope. COSMO-Onset is a Bayesian speech perception model associating a decoding module to process the spectro-temporal content of the speech input and a temporal control module enabling the segmentation of the speech input into constituent linguistic units. The decoding module has a hierarchical structure similar to classical psycholinguistic models like TRACE (McClelland & Elman, 1986), with three layers of representations (acoustic features, syllable, and word identity) usually considered in the context of isolated word recognition. The temporal control module associates a bottom-up mechanism for syllabic onset detection with an original top-down mechanism for syllabic onset prediction, involving temporal knowledge from higher linguistic levels (e.g., lexical or prosodic).

With this model, we explore the dynamics of speech segmentation resulting from the combination of such bottom-up and top-down temporal mechanisms. The fusion architecture has been developed in relation to the observed weaknesses of the existing neuro-computational models of speech perception. Hence, COSMO-Onset is developed to address the current limits of these purely bottom-up parsing systems.

Organization of the thesis

The present document is organized as follows:

Chapter 1, Oscillation-based models of speech perception We present a systematic review of the main neuro-computational models of speech perception by quickly providing an overview of neural oscillations and their relations to cognition in general, and to speech perception in particular. Then, we go through the main oscillatory-based speech perception models and present a key experimental paradigm to evaluate the potential role of top-down temporal prediction. We end

this chapter with a critical review of the literature which leads us to state the questions addressed in the current thesis.

Chapter 2, COSMO-Onset: The conceptual model We present COSMO-Onset, at the conceptual level, by describing the model architecture and its main interacting components, leaving out all specific details for further elaboration in the following chapters.

Chapter 3, COSMO-Onset: The illustrated model We present the first variant of the COSMO-Onset model, specifying all the components presented in **Chapter 2**. We then present the simulations on a representative set of "toy" stimuli and situations, elaborated for the purpose of illustrating the key concepts which allow assessing the role of the top-down onset prediction component.

Chapter 4, A study of the oscillation-based syllabic segmentation model by Räsänen et al. (2018) In order to deal with real speech stimuli, we present the oscillation-based model developed by Räsänen et al. (2018). We first evaluate it on a French corpus for syllabic onset detection and extend it to the detection of P-centers on the same French corpus. We then use it not only to assess whether isochrony plays a role or not in speech perception but also to evaluate its robustness to noise.

Chapter 5, A syllable recognition model using Random Forests Still, evolving towards a variant of COSMO-Onset able to deal with real speech, we present and evaluate a machine learning algorithm able to perform unit recognition.

Chapter 6, COSMO-Onset: Adapting to real speech We present an adaptation of COSMO-Onset for real speech input, capitalizing on the models/tools described in the last two previous chapters enabling us to deal with real speech. As in chapter 3, we first describe the model architecture and then present the main simulation results on a real speech corpus.

Chapter 7, Conclusion and Discussion We discuss the main findings of this thesis and the perspectives that it provides.

Chapter 1

Oscillation-based models of speech perception

The last decade has seen an increasing number of models of speech perception based on neural oscillations. This is a direct consequence of the fact that methods and evidence for the role of neural oscillations in speech perception are becoming increasingly ubiquitous.

At their core, all these models share the general assumption that the speech signal has some kind of oscillatory characteristics, and that brain neuronal populations would be able to track these oscillations in the acoustic signal, which would then hierarchically structure the speech decoding process over time.

In this section, we will review the literature on recent models of speech perception based on neural oscillations. We will start with two phenomenological/conceptual models, which are the two seminal works of Ghitza (2011) on the one hand and Giraud and Poeppel (2012) on the other. These will lay down the general principles at the core of every oscillatory model. Following Giraud & Poeppel's work, there have been two recent contributions that provide neurocomputational models, namely those by Hyafil, Fontolan, et al. (2015) on coupled cortical theta and gamma oscillations, and by Hovsepyan et al. (2020), which presents a decoding module strongly inspired at the architectural and technical levels by the hierarchy of nonlinear dynamical systems previously developed by Yildiz et al. (2013) for birdsong modeling. In parallel, it is interesting to consider the more functional model developed by Räsänen et al. (2018), which is based on linear signal processing techniques. All of these models bring together several fundamental principles highlighting the importance of neural oscillations in speech perception.

In this Chapter, we begin with a quick primer on neural oscillations in general. We then proceed to describe the natural rhythms of speech and their relation to neural oscillations, before reviewing the main oscillation-based models of speech

perception. Finally, we present an experimental paradigm that we consider as central for evaluating our proposed model and discuss more precisely the main assumptions and expectations of our work.

1 Neural oscillations: a quick overview

It is no secret that we are surrounded by many cyclic sequences of events. No huge effort is required to observe that the sun rises and goes down in a kind of regular manner, the church bell rings periodically, analog clocks tick at every second, the human heart beats at regular intervals (Clarke et al., 1976), plants have their own rhythms (Damineli et al., 2022), music is rhythmic, and its rhythm has often been compared to the rhythm in language (Bispham, 2006; Fiveash, Bedoin, et al., 2021; Hickok et al., 2015; Kotz et al., 2018; Patel & Daniele, 2003), etc. Interestingly, the human brain also features many rhythms. These rhythms are ubiquitous and are believed to play an important role in human cognitive processes, from memory, to attention, to thoughts, and even to consciousness (Baars & Gage, 2013; Buzsaki, 2006; Buzsáki & Draguhn, 2004; Cannon et al., 2014).

Simply put, neural oscillations are the rhythmic activity of ensembles of neurons in the brain. The coordinated electrical activity in a group of neurons gives rise to periodic or quasi-periodic rhythms in the brain. These electrical signals can be recorded using various invasive and non-invasive techniques, such as electrophysiology *in vitro* and *in vivo* (electroencephalography), optogenetics, and magnetoencephalography (Cannon et al., 2014; Keil & Senkowski, 2018; Mitra & Pesaran, 1999). Whether awake or at rest, the rhythms in the brain are a crucial part of its activity, and their history can be traced back to 1924 when Hans Berger discovered the human electroencephalography (Berger, 1929). Different types of neural oscillatory responses can be distinguished, each reflecting different aspects of neural synchronization (David et al., 2006; Tallon-Baudry & Bertrand, 1999). Evoked oscillations, for instance, are related to the presentation of an external stimulus. They are precisely phase-locked to this external stimulus and are displayed by averaging or summing different evoked oscillations over trials of identical phase; such analyses provide event-related potentials (ERPs) (Başar et al., 1999; Moratti et al., 2007). Induced oscillations, on the other hand, can occur independently of external stimulation. Induced oscillations represent local oscillations in a given range of frequency but without precise phasing to a reference time instant, internal or external, that is associated with the presentation of a stimulus (Klimesch et al., 1998).

Several studies have characterized neural oscillations. Even though it is still not fully established, a convenient classification of the rhythms has been adopted

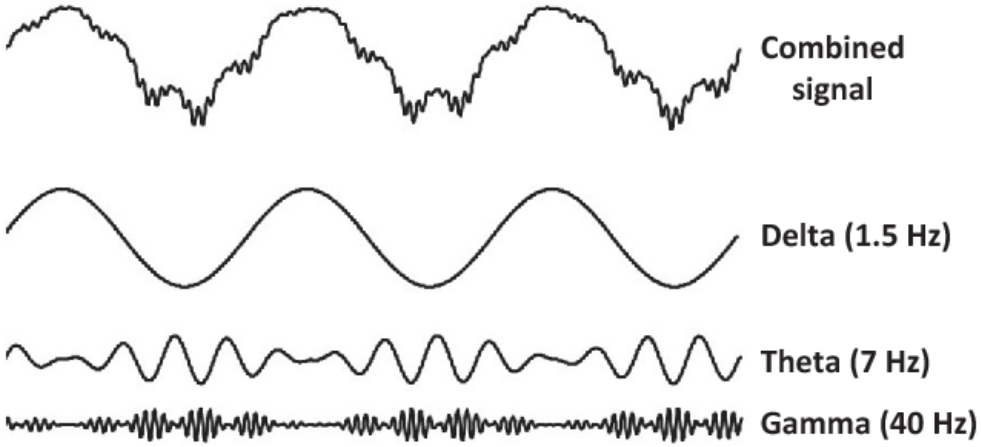

Figure 1.1: Example of a typical EEG signal and its decomposition into several oscillatory frequencies. On top, the full combined EEG signal, followed respectively by its delta frequency constituents, theta frequency constituents, and gamma frequency constituents. Image is taken from (Calderone et al., 2014).

with five gross frequency bands, namely the delta (δ, 1–3 Hz), theta (θ, 4–8 Hz), alpha (α, 9–14 Hz), beta (β, 15–30 Hz), and gamma (γ, $>$ 30 Hz) bands (Buzsaki, 2006; Clayton et al., 2015). Figure 1.1 shows a typical example of an electroencephalography signal (the top plot) with its decomposition into three oscillatory frequencies (the following three plots), respectively the δ (at 1.5 Hz), the θ (at 7 Hz) and the γ (at 40 Hz).

Various cognitive functions have been proposed for each of these characteristic bands (Amzica & Steriade, 1998; Başar et al., 2001; Başar et al., 1999; Bastos et al., 2012; Buzsáki & Draguhn, 2004; Fries, 2015; Lakatos et al., 2008; Schroeder & Lakatos, 2009). Thus, the delta band, which can be observed in sleeping cats and humans, has been associated to signal detection and decision making (Başar et al., 1999; Harmony, 2013; Nácher et al., 2013); the theta band has been associated with memory (Buzsáki, 1998; Reiner et al., 2014); the alpha band has been associated to attention and vigilance (Hanslmayr et al., 2011; Klimesch, 2012; Knyazev et al., 2004; Rohenkohl & Nobre, 2011); the beta band has been associated to motor planning Haegens et al. (2011), Morillon et al. (2019), and Tzagarakis et al. (2010); and finally, the gamma band has been associated to feature binding (Csibra et al., 2000; Fries, 2015; Singer, 2001). Although it is convenient to think that each neural oscillation band is associated with one or several specific cognitive processes, it is important to be aware that this view is likely simplistic. Several studies have shown that any of these frequency bands can be related to various mechanisms, and in different brain regions, depending on the cognitive process at hand (Ainsworth et al., 2011; Cannon et al., 2014).

2 The natural oscillations of speech dynamics

Speech is a signal with quasi-oscillatory properties, not only by its physical nature (Ding et al., 2017; Goswami, 2019; Poeppel & Assaneo, 2020), but also by its linguistic nature (Beckman, 1992; Cummins, 2015; Gibbon & Gut, 2001; Ramus et al., 1999). The quasi-rhythmic nature of the speech signal likely comes from the physical dynamical properties of the articulators driving its content, particularly from the jaw, providing natural syllabic "frames" embedding the other articulatory movements according to the frame-content theory of evolution of speech production (MacNeilage, 1998; MacNeilage & Davis, 2000). More globally, studying its waveform suggests a potential hierarchy of embedded amplitude or spectral modulations that can be related to the different linguistic units. A first basic amplitude modulation rhythm, likely related to jaw dynamics as mentioned previously, is provided by the syllabic rhythm with an average syllable duration roughly estimated at 250 ms (Greenberg, 1998; Greenberg et al., 2003), corresponding to a frequency around 4 Hz and related to the theta frequency band (4 to 8 Hz) in neural responses. Within the syllabic frame containing a variable number of phonemes, it has been proposed that minimal decoding units related to acoustic phones could have an average duration roughly estimated at 25 ms, corresponding to a frequency of 40 Hz, related to the gamma frequency band ($>$ 30 Hz) for phoneme analysis. At a larger temporal scale, slower delta band oscillations (1-3 Hz) are related to syllable sequences and words embedded within prosodic phrases, with duration roughly varying in the interval of 500–2000 ms. Figure 1.2 shows an example of speech input and its different linguistic contents, which are hierarchically embedded within each other (higher linguistic units contain lower linguistic units). These characteristics may be universal; at least, they are present in many languages of the world (Ding et al., 2017; Fiveash, Falk, et al., 2021; Luo & Poeppel, 2007; Pellegrino et al., 2011). It thus has been hypothesized, in light of these ubiquitous relations between brain and speech rhythms, that brain oscillations may play a crucial role in speech perception (Ahissar & Ahissar, 2005; Arnal, 2012; Arnal et al., 2015; G. J. Brown et al., 1996; Chandrasekaran et al., 2009; Ghitza & Greenberg, 2009; Giraud et al., 2007; Giraud & Poeppel, 2012; Kösem et al., 2018; Kösem & Van Wassenhove, 2017; Peelle & Davis, 2012; Poeppel & Assaneo, 2020).

Importantly, slow oscillations in the delta and theta frequency bands might be useful in the temporal organization of speech at various levels, respectively in syntactic parsing and syllabic chunking. Higher frequency oscillations in the gamma frequency band might be related to decoding the spectral information of the speech. Finally, the coupling between slow and high oscillations would give rise to the neural mechanisms underlying speech perception, as in many other

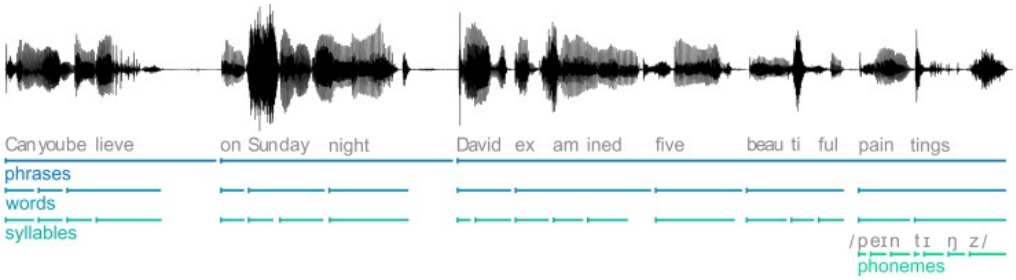

Figure 1.2: Example of speech input divided into its linguistic constituents at different time scales. On top, the speech waveform of the sentence "Can you believe on Sunday night, David examined five beautiful paintings". Following respectively, the phrasal segmentation, the words segmentation, the syllabic segmentation, and a phonemic segmentation for the last word of the sentence "paintings". Image is taken from (Keitel et al., 2018).

cognitive functions such as vision (Calderone et al., 2014; Canolty et al., 2006; Canolty & Knight, 2010; Fries, 2015; Ward, 2003; Wyart et al., 2012). It has been shown that the phases of the slow oscillations are much related to high oscillations powers. Typically, the delta phase is coupled to theta amplitudes, in a way that the theta amplitude is larger during one phase of the delta and smaller during the opposite phase. Similarly, theta phase is coupled to gamma amplitude. The coupling mechanism between these different neural oscillations can be seen on Figure 1.1.

The interaction between lower and higher processing stages is well within the predictive coding realm (Friston, 2005; Friston & Kiebel, 2009; Mumford, 1992; Rao & Ballard, 1999), in which the brain is considered to be a "predictive machine" where internal states support inferences about sensory stimuli. Importantly, neural oscillations in the beta band have been hypothesized to be a preferential channel for the exchange of information from the higher to the lower layers of processing (Arnal & Giraud, 2012; Engel & Fries, 2010; Rimmele et al., 2018; Sohoglu et al., 2012).

3 The Dynamic Attending Theory and its implications for speech

Almost all ecological stimuli can exhibit some degree of temporal regularity. For M. R. Jones (1976), early psychology researchers did not pay enough attention to the importance of time as a sensory dimension. The temporal context in which physical events occur has an influence on how those same events are perceived. She proposes a general framework, the **Dynamic Attending Theory (DAT)** which puts a special emphasis on the temporal aspect of auditory perception, memory, through attention (M. R. Jones, 1976; M. R. Jones & Boltz, 1989). In

her seminal paper, M. R. Jones (1976) proposes a general theory of perception based on two fundamental aspects.

The first is what the author called the "Subjective Representations of the Physical World", involving four (4) assumptions:

1. the "physical dimensions" assumption: patterns of the world are best defined by three spatial dimensions, completed by a time dimension. Importantly, as the author puts it: "All dimensions are most simply conceived as having nested, or hierarchical, structure";

2. the "invariance" assumption: even though the physical dimensions can change, the hierarchical structure of world patterns is explained in terms of invariant relations;

3. the "subjective dimensions" assumption: we can somehow assign a subjective counterpart to each physical dimension;

4. the "subjective pattern structure" assumption: relations that characterize subjective pattern structure can be expressed in terms of subjective dimensions.

The second aspect of the theory concerns the "interaction of organisms with the real world", with a fundamental premise, assuming organisms have their own rhythms which are more or less related to the perceptual rhythms of the physical objects through their temporal dimension. The first two assumptions ("rhythmic organisms" and "synchrony"), although motivated by behavioral results, have an apparent link with neural oscillations, since they state that, intrinsically, there are biological rhythms, and that these rhythms have perceptual equivalents, leading to a phenomenon called "entrainment" which enables the tracking of expected events.

From these bases, the **Dynamic Attending Theory**, as the name suggests, is a theory about attention, where the fundamental theoretical proposal is that attention is not static, that is, constant over time, but rather dynamical, that is, varying over time. This dynamical nature of attention is related to the temporal dimension of ecological stimuli, which adds on top of their usual spatial dimensions; for instance for visual objects, we have the width, height, and depth, whereas, for auditory stimuli, we have the subjective dimensions of pitch and loudness. The "entrainment" hypothesis states that internal rhythms are driven by and synchronize with the external rhythms present in stimuli. Large and Jones (1999) proposed a model using two fundamental aspects: "self-sustaining oscillations" and "energy pulse". The first aspect ("self-sustaining oscillations") generates temporal expectancies which are periodic in the absence of external stimuli, but when coupled with an external stimulus, gets "synchronized" or

"entrained". In contrast, the second aspect ("energy pulse"), modeled by a probability distribution, is concerned by the attention dynamics involved in model predictions enabling to take into account natural rhythmic fluctuations of attention.

This simple concept is supported by a multitude of empirical data, which show that the detection and the discrimination of target stimulus events, together with response times to target stimulus events, vary depending on how the target events are related to their temporal context, in other words how aligned are both the target events and the attending rhythm.

The **Dynamic Attending Theory** happened to provide a perfect cognitive psychology conceptual framework for all the findings and proposals that emerged in the last 30 years about the role of resonant and oscillatory phenomena in the human brain. Focusing on audition and the processing of acoustic stimuli, it has been widely explored and tested in the field of music perception (e.g., M. R. Jones & Boltz, 1989; M. R. Jones et al., 2002; Tillmann et al., 2006; Tillmann & Lebrun-Guillaud, 2006). In relation to speech perception, a number of studies, inspired by the DAT, explored the role of rhythmic cueing and/or rhythmic priming on phonological and syntactical processing, speech production and interaction, and the remediation of auditory and speech handicaps (e.g., Cason & Schön, 2012; Hidalgo et al., 2019; Kaya & Henry, 2022; Obleser & Kayser, 2019; Quené & Port, 2005; Schön & Tillmann, 2015). They all converged on the fact that it is possible to enhance a listener's or speaker's performance in various aspects of perception or production of speech by focusing her attention in time on the regulation of specific speech events, possibly through the preliminary exposition to rhythmic stimuli capturing some of the properties of the corresponding speech events. Even though the Dynamic Attending Theory will not be further mentioned in the present document–and is seldom mentioned in neuroscience papers–it provides an underlying pivot of important concepts that will be exploited all along this work.

4 Conceptual oscillation-based models of speech perception

On this global neuro-behavioral basis, two conceptual models emerged in the years 2010, which paved the way for most further developments in the field.

4.1 The *TEMPO* model by Ghitza

4.1.1 Model description

Figure 1.3 shows the most recent version (Ghitza, 2020) of the ***TEMPO*** architecture initially proposed by Ghitza (2011). In this model, the speech input first goes

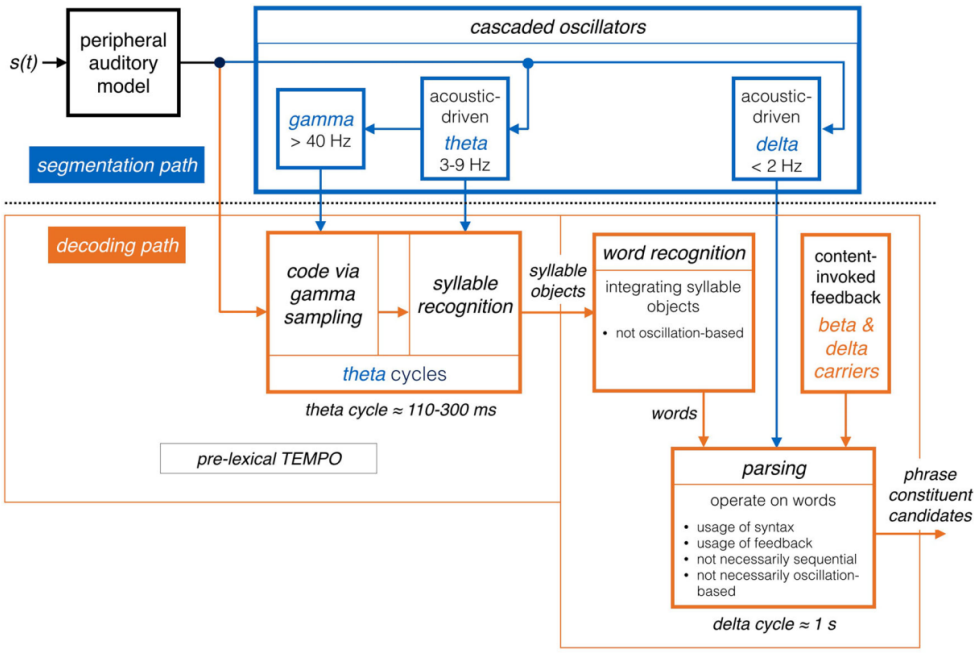

Figure 1.3: The most recent architecture of the ***TEMPO*** model. The segmentation path is the upper part (in blue) with the cascaded oscillators, namely the delta oscillator to provide temporal frames for words-phrase analysis, the theta oscillator to provide temporal frames for syllable parsing, and the gamma oscillator to provide neural codes for later pattern matching. The decoding path is the lower part (in orange) where pattern matching and recognition are performed within the different temporal frames controlled by the segmentation path, from the phonetic constituent to the phrasal constituents, passing through the syllabic constituents. Image is taken from (Ghitza, 2020).

through a peripheral auditory model, for basic spectro-temporal analysis. From there on, the ***TEMPO*** architecture contains two information processing channels (Figure 1.3, in blue and orange, respectively). The first path, the "segmentation path", is dedicated to processing the temporal aspects of speech. It is organized as a hierarchy of oscillators in which high-frequency oscillators are embedded in low-frequency oscillators. In this set of cascaded oscillators, the central one, in the theta frequency band, is the pivot oscillator, which tracks the syllable rhythm in speech. It monitors the higher-frequency gamma oscillator providing spectro-temporal information at an infra-syllabic rate above 40 Hz, and it is under the temporal control of the lower-frequency delta oscillator driving phrasal fluctuations in the 1–3 Hz range. This path would therefore provide temporal frames in which to decode the acoustic content in order to perform pattern matching with syllabic patterns stored in the memory. This pattern matching is performed in the second path, the "decoding path", concerned with decoding the speech content. In a theta oscillation cycle, there would be a fixed number

of high-frequency oscillations in order to organize the processing of sub-lexical phonetic units such as phones.

A fundamental characteristic of the coordinated operation of these oscillators is related to the fact that they are able to adapt their rhythm to the rhythm of the speech input, thus facilitating the integration of prosodic variations in the signal. It should therefore be noted that these oscillators are not totally resonant systems that would stay perfectly periodic, but rather gently resonating input-driven systems that oscillate in a pseudo-periodic way, with the constraint that each oscillator has a fixed preferential frequency band.

4.1.2 Main results

Although there is no complete implementation of ***TEMPO***, Ghitza has shown qualitatively that the model is capable of simulating several sets of behavioral data, that are claimed to be beyond the scope of conventional speech perception models, which only feature an acoustic signal decoding channel. These include data from Ghitza and Greenberg (2009) on the intelligibility of compressed speech. Since this study is influential, let us describe it in some more detail.

One of the main sources of variability in speech is due to how fast or slow some talkers speak. In both cases, this may result in distorted speech that can prove difficult for listeners. Usually, in ecological communication, listeners seldom face slowed speech, but are more confronted with accelerated speech, which can be thought of as "compressed speech". Such alterations of the speech rate have consequences on information rate at various levels, namely at the prosodic and acoustic levels (Arons, 1992; Foulke, 1971; Foulke & Sticht, 1969; Maki & Beasley, 1976). Since speech-altered rhythms differ from the "canonical" speech rhythms listeners expect, one important question is how they cope with it and whether there is any relation between time-compressed speech perception and neural oscillations. Ghitza and Greenberg (2009) address this question by experimentally inserting silence segments in compressed speech, and studying how this would help recover intelligibility.

There has been research in the literature, prior to these, investigating the temporal mechanisms at play in speech perception. A noteworthy study by Huggins (1975) investigated the hidden temporal variables that underlie intelligibility. Contrary to Huggins (1975), who "suggested that the factor governing intelligibility was not phonetic glimpsing per se, but rather some internal time constraint on processing spoken material" tightly related to an echoic memory buffer, Ghitza and Greenberg (2009) suggest that "the decline in intelligibility is the result of a disruption in the syllabic rhythm beyond the limits of what brain neural circuitry can handle". In other words, they argue for a potential role of neural oscillations, underpinning the intelligibility of compressed speech.

Ghitza and Greenberg (2009) used the SUSGEN corpus (Bunnell et al., 2005) which contains short (of about 2 seconds in duration) Semantically Unpredictable Sentences (SUS). They first showed that if compression is too strong (e.g., synthetically accelerating the rhythm by a factor 3) intelligibility of such unpredictable sentences dropped severely (with up to 50 % words not recognized), but it could be largely improved by inserting silence gaps, decreasing word error rate down to 20 % in the best cases. The authors claim that inserting silence gaps provides a way to align acoustic information between the compressed and uncompressed signal, thus allowing the recovery of potential information loss. They link this result with neural oscillations tracking mechanisms, suggesting that silence gaps insertion constrains the decoding of the speech input in an interval of time that is suitable to the endogenous brain rhythms involved in speech processing, namely the theta band.

Conventional models with only the decoding channel are unable to simulate such experimental observations, whereas with the ***TEMPO*** model, thanks to the pseudo-periodic and adaptive oscillators, it is possible to simulate them. The moments of silence play the role of calibrating the temporal decoding frames necessary to decode the different acoustic chunks, which are in the frequency band corresponding to the syllabic rhythm (Ghitza, 2013).

We can hereby note that, in cases of ecological communication, where speech is not compressed nor degraded, the ***TEMPO*** model would operate in a way similar to the classical models, with less evidence in favor of the crucial role played by the cascaded oscillators. On the other hand, in order to explain results such as those presented by Ghitza and Greenberg (2009), the segmentation path appears to be critical.

4.2 The model by Giraud & Poeppel

4.2.1 Model description

Figure 1.4 shows the conceptual model of speech perception proposed by Giraud and Poeppel (2012). Unlike Ghitza's ***TEMPO*** model, where we can clearly distinguish between two processing streams, here the authors proposed a step-by-step data processing mechanism applied sequentially to the speech signal. Importantly, at the output of the spectro-temporal analysis stage in the cochlea, the signal is converted in a spike train which then constitutes the input to all further processes.

The first step is called the **"Phase reset"** step. At rest, intrinsic neural oscillations are already present in the cortical activity of the human brain, and this happens in various bands of oscillations (Deco & Corbetta, 2011; Giraud et al., 2007; Laufs, 2008; Northoff et al., 2010). In this first step, the intrinsic

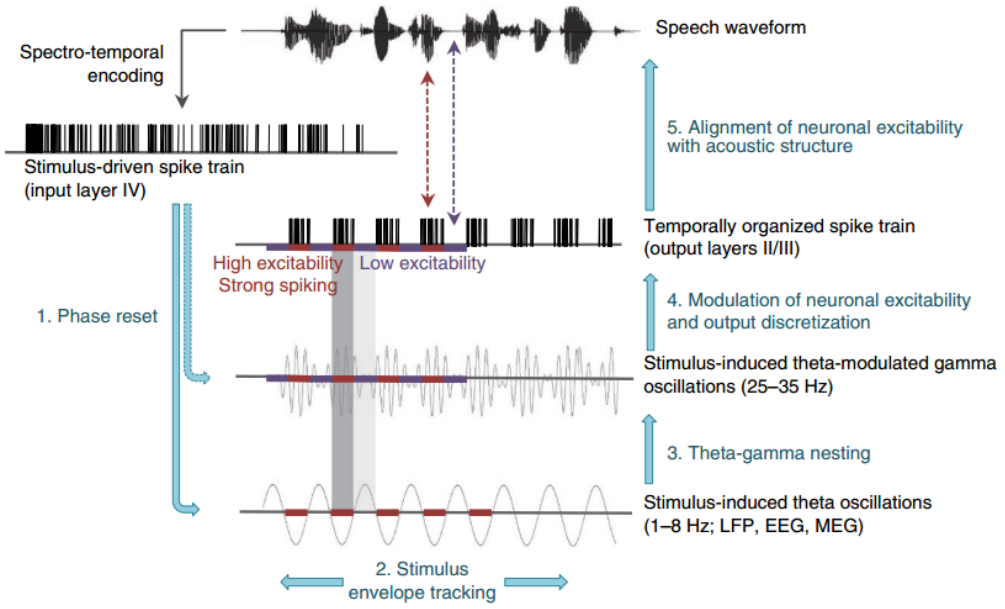

Figure 1.4: The conceptual model of speech perception by Giraud and Poeppel (2012). The speech waveform goes through hierarchical processing from higher cortical layers to lower cortical layers (Input layer IV, output layers II/III). The communication between these layers is achieved through the synchronization of neural oscillations (stimulus-induced ones, local field potentials (LFP), recorded by electroencephalography (EEG) or magnetoencephalography (MEG)). Image is taken from (Giraud & Poeppel, 2012).

oscillations are reset to align with the proper dynamics of the input data, in some specific bands, notably in the theta band, but not only.

These specific oscillations track and entrain to the speech input stimuli in the 4–8 Hz theta frequency band, facilitating the tracking of salient points present in the speech envelope. This is the second step, also called **"stimulus envelope tracking"**.

Then in step 3, called **"Theta-gamma nesting"**, with the theta oscillations orchestrating the gamma oscillations, the lower frequency oscillations (theta) organize the higher frequency oscillations (low gamma, in the 25–35 Hz frequency band). This allows the processing of phonemic units with lower time scales in the gamma band, inside larger syllabic units within the theta band. It is still a matter of debate how many high-frequency cycles are processed within a full cycle of a lower-frequency band. This cross-frequency organization ensures and organizes successful fine-grained processing of the speech input.

Then the activity in the low and high gamma oscillatory bands modulates the spike trains encoded from the input signal. This is the fourth step, also called **"modulation of neuronal excitability and output discretization"**.

Finally, the temporally organized spike train can be used to search for codes

of discrete linguistic units, especially at the syllabic and phonemic levels. This is the fifth step, also called **"alignment of neuronal excitability with acoustic structure"**. How this mapping is precisely performed is not described in detail.

While the model proposed by Giraud and Poeppel (2012) is rather conceptual, there have been several developments towards more or less realistic neuro-computational models based on these principles, which we describe later on. The first one developed by Hyafil, Fontolan, et al. (2015), is described in section 5.3. The most recent one by Hovsepyan et al. (2020) is presented in section 5.5.

4.3 General principles of oscillation-based models

The two models by Ghitza (2011) and Giraud and Poeppel (2012) allow us to establish a number of general principles that would underlie all models of speech perception based on neural oscillations.

The first fundamental principle concerns the dissociation of the information processing pathways by clearly separating the temporal mechanisms from the spectral content decoding mechanisms. As in classical models, we still have a spectro-temporal decoding pathway, but this is guided by a temporal segmentation pathway that provides potential temporal linguistic boundaries in the signal. Even if the separation does not clearly appear in Giraud and Poeppel (2012), in ***TEMPO***, it is clear that the architecture is designed so as to put an emphasis on the separation of the temporal and decoding mechanisms.

The second principle refers to the nesting of oscillations in different frequency bands, with the higher frequency oscillations embedded in the lower frequency ones. How many cycles of high-frequency oscillations are embedded within cycles of low-frequency oscillations is still a partly open question that has generated various proposals. In the first developments of the ***TEMPO*** architecture (e.g., Ghitza, 2013), Ghitza suggests that in a cycle of theta oscillations, one could expect typically 4 cycles of beta oscillations with similar duration, as well as 4 cycles of gamma oscillations in each of the cycles of beta oscillations, thus making 16 cycles altogether in a theta oscillation. In contrast, Giraud and Poeppel (2012) propose that there would be at most 4 cycles of gamma oscillations in a cycle of theta oscillation – although later proposals by the same group enacted other choices (for example, Hovsepyan et al. (2020) used 8 gamma oscillations framed in a cycle of theta oscillation).

5 Computational oscillation-based models of speech perception

Based on these principles, several neuro-computational models have been proposed and tested on different tasks, such as word or syllable recognition tasks.

Although conceptual models summarize the general principles underlying speech perception based on neural oscillations, there remains a gap to fill when it comes to testing the validity of their predictions and hypotheses on real speech stimuli. Here, we will describe the main models implementing neural oscillations in order to segment speech, in particular into syllable-like acoustic chunks. For each model, we will describe how it works, and then we will present its main evaluation results. As the performance metrics are almost the same for all these models, we will summarize them before starting the review of the models per se.

5.1 Metrics to evaluate models

Most often, in evaluating computational models of speech perception, one might be interested either in the model performance in correctly categorizing the speech input ("unit performance"), or its temporal accuracy in relation to the syllabic/lexical boundaries of the speech input ("boundary performance"). Depending on the task, in the first case, one assesses how well the model recognizes speech units, which might be the phoneme, the syllable, the word, or a higher-level structure such as the whole sentence. In contrast, in the second case, regardless of the considered unit, one evaluates the correctness of the model's temporal predictions.

These metrics are computed by comparing with human-annotated data or semi-automatically annotated data with human correction thereafter. Such data are often called "ground truth", as they provide the "real" boundaries and units to detect and categorize. Thus, the evaluation of computational speech segmentation models would boil down to comparing the outputs of the models with the corresponding ground truth.

In unit identity recognition, we often ignore temporal alignment details. The task is about the identification of a lexical unit (e.g., syllable or word) or a sequence of lexical units, and we only compare the model's output with the real lexical units to identify. Whatever the case, we can consider that the model outputs a sequence of predicted units (a sequence of only one unit in the case of isolated unit recognition), that have to be compared to a reference sequence of labels. The evaluation then boils down to a comparison of sequence matching between the predicted and the reference. Classically, errors of prediction can be identified either as a substitution error (S), a deletion error (D), or an insertion error (I). Two metrics can then be computed based on these errors: the percentage correct, PC and the percentage accuracy, PA (Young et al., 2002), the difference being that the first ignores the insertion errors whereas the second penalize them. The percentage correct is defined as follows:

$$PC = 1 - \frac{D + S}{N},$$

with N the total number of true units to be recognized. And the percentage accuracy is defined by:

$$PA = 1 - \frac{D + S + I}{N},$$

where the values of *D, S, I* can be computed using any sequence matching algorithm, for example, the one based on the Levenshtein distance (Young et al., 2002; Yujian & Bo, 2007).

Temporal segmentation involves the analysis of model temporal predictions compared with a given ground truth. The ground truth can either be syllabic onsets (in other words, instants of syllable beginnings) or syllable P-centers (usually defined as the psychological moment of occurrence of syllables Morton et al., 1976). The syllable onsets are considered to correspond to troughs of the envelope of the speech signal, whereas P-centers correspond to peaks of the derivatives of the speech signal envelope (Marcus, 1981; Patel et al., 1999). Because of possible annotation errors, coming either from humans, from the tools used, or from elsewhere, authors often allow a margin of error around the proposed ground truth events, typically 50 ms (Obin et al., 2013; Villing et al., 2006). Performance metrics involve model precision, or recall, or its F-score, which is a combination of the two (Chinchor, 1992; Sasaki et al., 2007). The precision P is defined as the proportion of events predicted by the model that corresponds to real events, the recall R as the proportion of real events correctly predicted by the model, and the F-score is a combined measure that performs a trade-off between the two, and is calculated using their harmonic mean:

$$F = \frac{2PR}{P + R}.$$

These metrics are used in many problems dealing with evaluating model performance such as in event detection or sound detection tasks (Heittola et al., 2013; Temko et al., 2009).

Since these two metrics evaluate different aspects of the model's performance, it is possible to combine them into aggregate measures that calculate the overlap of model predictions and ground truth. One can find variants such as the temporal ratio of correct lexical categorization (Hovsepyan et al., 2020), or the unit recognition accuracy within the correct intervals (Räsänen et al., 2018). Both variants consider both categorization and event detection performance. The more a model predicts a correct category in a time interval that overlaps as much as possible with the ground truth temporal interval of this category, the better the model is considered to be.

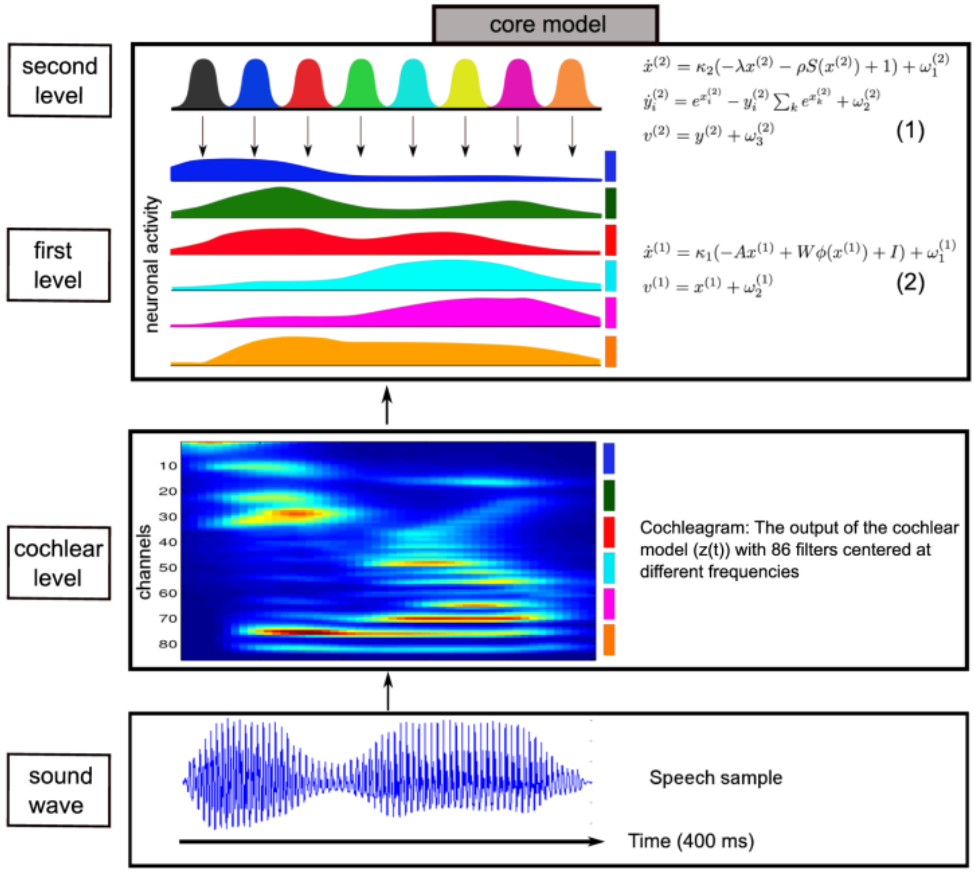

Figure 1.5: Yildiz et al. (2013)'s model of speech perception inspired by a birdsong recognition model. Image is taken from (Yildiz et al., 2013).

5.2 The model by Yildiz, Kriegstein & Kiebel (2013)

5.2.1 Model description

Figure 1.5 shows the model developed by Yildiz et al. (2013). It is highly inspired by the hierarchical model that they developed for birdsong recognition, which is based on the accumulated knowledge of how birds perceive songs (Yildiz & Kiebel, 2011). The rationale relies on first acknowledging that speech and birdsong are both, first and foremost, sound waves; second and most importantly, there also are various similarities in how birds process sounds and how humans process speech (Berwick et al., 2012; Bolhuis et al., 2010; R. Dooling, 1992; R. J. Dooling et al., 2002; Doupe & Kuhl, 1999). From there on, Yildiz et al. (2013) design a word recognition model based on the key principles of their previous song recognition system, that exploits Bayesian inference within dynamic hierarchical generative models, each associated with one of the linguistics units to be recognized, and more precisely, one per word.

The speech sound wave is first processed in a cochlear filtering model providing

a cochleagram with 86 frequency channels, which is then used as the input to the two-level hierarchical core model. The cochleagram output is down-sampled in order to finally keep only 6 dimensions by linear window averaging every consecutive 14 channels (removing the last two channels of highest frequency).

The 6-dimensional input is then fed into a series of generative modules, one per each word to be recognized. Generative modules combine a higher level ("second level" in Figure 1.5) which controls in time a sequence of local states within each of the 6 spectral channels ("first level" in Figure 1.5). More precisely, the output of the first level represents the typical neuronal activity for a given word in a given frequency band, estimated by using a Hopfield attractor model (Hopfield, 1982; Hopfield & Tank, 1985). At the second level, the model interacts with the output from the first level through the activity of various neuronal ensembles and generates a new encoding of the speech input, which is now temporally organized exhibiting a sequential activation of neural sets following a winnerless competition principle, that is, a dynamical principle of brain dynamics where different neuronal ensembles change states sequentially dependent on the stimulus (Afraimovich et al., 2004; Rabinovich et al., 2001; Seliger et al., 2003). In their implementation, the authors propose to use a constant sequence of 8 successive states to describe each word in the corpus.

Therefore, globally, the first level of the core model results in a content decoding module, controlled in time by the second level which temporally organizes the outputs in order to guide the search for lexical units. Each word generative module is then fed with the speech input, resulting in the output in a "prediction error" which can be compared with the errors from all the other word generative modules. Categorization follows by searching the minimal predictive error in output in a Bayesian inference process which combines prior (top-down) information from the word generators previously learned from a learning corpus, with the incoming (bottom-up) acoustic data. The combination can be controlled from a so-called "precision" variable which enables modulating the relative importance of top-down and bottom-up information in the decision process.

5.2.2 Main results

To demonstrate the behavior of their model, the authors performed a word learning and recognition task, using a dataset considering digits from zero to nine (Instruments, 1991). Overall, authors report a Word Error Rate (WER) of 1.6 %, which is, according to them, on par with state-of-the-art automatic speech recognition models for the same kind of material (isolated digits) in the literature.

Moreover, in order to account for ecological speech communication situations that human listeners face daily, they tested the model in various "degraded"

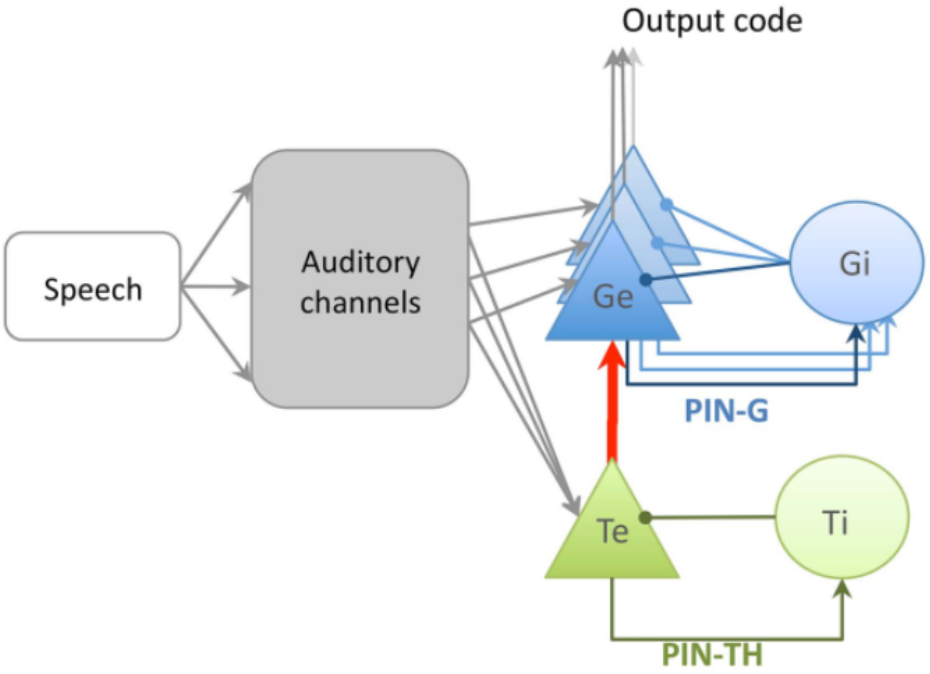

Figure 1.6: Hyafil, Fontolan, et al. (2015)'s model of speech perception. The Pyramidal Inter Neuron- Theta (PIN-TH) is composed of excitatory theta neurons (Te) and inhibitory theta neurons (Ti). Whereas the Pyramidal Inter Neuron-Gamma (PIN-G) is composed of excitatory gamma neurons (Ge) and inhibitory gamma neurons (Gi). Image is taken from (Hyafil, Fontolan, et al., 2015).

scenarios. First, they showed that the model was robust to varying levels of noise, with the worst WER of 11.2 % which compares, once again, to the state-of-the-art in similar conditions. Second, they evaluated the model performance in compressed speech with varying speech rates. They showed that the model was still able to perform quite well up to 25 % compression rate. Authors attribute this result to the fact that the model is able to adapt to the rhythm of the signal, and that even for time-compressed speech, the sequence of dynamics is less affected by the temporal change overall in the input.

5.3 The model by Hyafil, Fontolan, Kabdedon, Gutkin & Giraud (2015)

5.3.1 Model description

Figure 1.6 shows the model developed by Hyafil, Fontolan, et al. (2015). It is a neuronal spiking model based on the cross-frequency coupling between theta oscillations and gamma oscillations. The main focus is to precisely describe

the oscillations that are observed in human cortical activity. The model takes inspiration from a model implementing the neural oscillations in the gamma band, namely the Pyramidal Inter Neuron Gamma (PING) which simulates the spiking rhythm by alternating bursts of inhibitory neurons (Gi cells) and bursts of excitatory neurons (Ge cells) (Ainsworth et al., 2011; Jadi & Sejnowski, 2014). In their model, Hyafil, Fontolan, et al. (2015) assumed the same mechanisms of oscillation generation for both gamma and theta neural oscillations, naming the latter Pyramidal Inter Neuron Theta (PINTH), with Ti and Te cells, respectively for inhibitory and excitatory theta neuronal populations. Neurons for both PING and PINTH populations were modeled using leaky integrate-and-fire neurons (Börgers et al., 2005; Börgers & Kopell, 2008; Burkitt, 2006).

The two neural oscillations resonate at different time scales: PING at the gamma timescale (25–40 Hz), and PINTH at the theta timescale (4–8 Hz). In the absence of stimulation, they both sustain activity at their respective intrinsic rhythms. In the presence of speech input, they both re-organize their activity accordingly. The theta neuronal population activity (PINTH) is first simulated by all the auditory channels of the speech input. It then locks to the speech input in its specific rhythmic range by tracking slow amplitude modulations. On the contrary, the gamma activity (PING) with its different excitatory modules, is simulated by specific auditory channels. It fires at rapid rhythms tuning into the fine details of phonemes. Following the presentation of speech, the PINTH activity, which entrains to the envelope, constrains the PING activity whose outputs can be used to recognize the syllables. The authors designed the PINTH carefully in such a way that it follows the syllabic rhythm in the theta neural oscillations 4–H Hz range. It is thus used to detect syllabic boundaries in the presented speech stimuli. This information can then be used to temporally organize the PING outputs into syllable units.

5.3.2 Main results

In order to evaluate model performance, Hyafil, Fontolan, et al. (2015) first compared their model to two other models of the literature in a syllable alignment task: a simple linear-nonlinear acoustic boundary detector based on a generalized linear point process model and a state-of-the-art off-line model developed by Mermelstein (1975), which simply identifies the local minima in the speech envelope as syllable boundaries. They used the speech corpus from TIMIT (Garofolo, 1993), which contains phonetically labeled English sentences. The results they obtained show that their model outperforms other models for both natural speed and compressed speech, with their model being the only one able to adapt to speech compression (Hyafil, Fontolan, et al., 2015, Fig. 2).

Hyafil, Fontolan, et al. (2015) also evaluated their model's performance by

comparing it with two variants, in an "artificial lesioning" model comparison study. From the main model, which they called the "Intact model", they considered two variant models: one where the theta activity was not driven by the speech stimulus, called the "Undriven model", and another where there was no coupling between theta and gamma oscillations, called the "Uncoupled theta/gamma model". Before evaluating model performance on real speech, they used simple temporal stimuli for two different tasks: stimulus classification measuring what we call the "unit performance" in section 5.1, to assess model categorization performance (Hyafil, Fontolan, et al., 2015, Fig. 3C), and stimulus detection measuring what we call "boundary performance" in section 5.1, to assess model performance at detecting the temporal boundaries (Hyafil, Fontolan, et al., 2015, Fig. 3D). They showed that the Intact model outperformed the other two models when combining the false alarm and hit rates; the Uncoupled model had the second-best performance, suggesting the importance of having the theta module driven by the speech stimulus. This can be explained by the fact that in the case where the theta oscillations are stimulated by the input, there would be an adaptation of the intrinsic rhythm to follow the one present in the signal, while in the other configuration ("Uncoupled theta/gamma model"), this would not be the case.

Finally, they assessed the model's performance on real speech decoding by considering the generated spike patterns as codes to recognize. Here also, as expected, the Intact model outperformed the other models, by achieving a performance of correctly classified syllables of 58 % (in a set of 10 possible randomly chosen syllables). It is noteworthy that, when considering spike counts instead of spike patterns as codes to recognize, the model performance was decreased by about 10 points, suggesting that spike patterns represent a better distinguishable neural code for categorization.

5.4 *RDF*: The model by Räsänen, Doyle & Frank (2018)

Note

The text of this section is partially adapted from the paper (Nabé, Diard, et al., 2022).

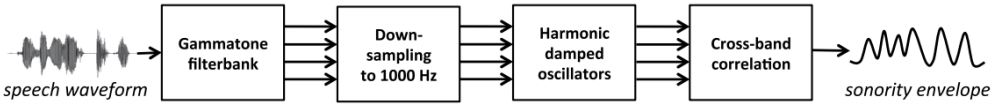

Figure 1.7: Räsänen et al. (2018)'s model of speech perception. Image taken from (Räsänen et al., 2018).

5.4.1 Model description

Figure 1.7 shows the architecture of the model developed by Räsänen et al. (2018), that we refer to as the ***RDF*** model hereafter. The authors developed an oscillator model for sonority-based rhythmic segmentation to study the pre-linguistic segmentation of syllables. It is mainly focused on the key mechanisms involved in language acquisition in infants (who supposedly have neither phonological nor lexical knowledge). The ***RDF*** model is on par with other oscillatory models of speech perception since they are all grounded on the same general principles, with a system driven by the energy fluctuations present in the speech signal (Arnal, 2012; Ding & Simon, 2014; Ghitza, 2011; Giraud & Poeppel, 2012). The model is based on the concept of sonority, that is the principle, elaborated in phonological theories of syllabic structure, that syllables are organized along sequences of minima and maxima of acoustic energy, minima corresponding to syllable onsets and offsets, and maxima to the vocal climax. Despite the debate around the physical reality of sonority, or whether it contributes to speech understanding (Daland et al., 2011; Harris, 2006; Parker, 2012), several studies have shown that sonority is correlated to syllabicity, namely with the Sonority Sequencing Principle (Clements, 1990), and that it is predominant in how infants perceive speech (Gómez et al., 2014; Hamza et al., 2018; Maïonchi-Pino et al., 2012; Price, 1980).

The ***RDF*** model of speech perception, based on linear second-order oscillators following cochlear analysis, is attractive since it is in some sense the simplest model that could be proposed to analyze the role of oscillations in processing envelope modulations for syllabic boundary detection. Starting from the speech signal, a set of signal processing techniques are applied in order to obtain an estimate of the sonority of the signal, as is depicted in the block diagram of the model (see Figure 1.7). First, Gammatone filter-banks (Holdsworth et al., 1988; Patterson et al., 1987) are applied to the speech input to get the amplitude envelope in 20 logarithmically spaced frequency bands. Their outputs are low-pass filtered and down-sampled to have an overall sampling rate of 1,000 Hz. Each envelope of each frequency band is then passed to a harmonic oscillator which resonates at a central frequency f_0 within a bandwidth Δf. Together, these define the oscillator Q factor, $Q = f_0/\Delta f$. Finally, the N most energetic outputs (usually N taken between 6 and 16) are combined by taking the sum of the logarithms of the amplitudes to obtain the sonority output; the final values are normalized between 0 and 1 over the stimulus duration. The resulting sonority function, which is a good estimate of the slow amplitude modulation, can be used in various ways, to identify speech-relevant events. The authors used it, in particular, to detect syllable boundaries by identifying local minima (valleys) in the sonority output.

5.4.2 Main results

The authors evaluated the model on a syllabic segmentation task using three publicly available speech corpora respectively in three languages with male and female speakers: the Switchboard corpus of spontaneous telephone conversations in American English (Godfrey et al., 1992), the phonetic corpus of Estonian spontaneous speech (Lippus et al., 2013), and the FinDialogue corpus of spontaneous Finnish speech (Lennes, 2009). All three corpora were annotated at the syllable level, with annotations verified by humans; this was considered the target ground truth. Model performance was compared to three syllabification algorithms of the literature: an envelope velocity-based algorithm proposed by Villing et al. (2004), a simple amplitude envelope minima detector using ear-like temporal filtering, and another sonority-based speech rhythm estimator developed by Wang and Narayanan (2007).

Four evaluation metrics were used. The first one assessed the temporal accuracy of syllable segmentation, by verifying whether predicted syllable boundaries fell with a margin of 50 ms, sooner or later than ground truth syllable boundaries ("boundary performance"). The second one assessed syllable temporal extraction, by considering not only the correct detection of the boundaries but also the detection of the beginning and end of syllables ("unit performance"), still with an acceptation margin of 50 ms. Then these two metrics were applied to words instead of syllables, that is, detection of word boundaries and word temporal extraction, with the same 50 ms precision criterion.

The oscillator model performed better than the other three models overall in both the boundary detection and unit detection tasks. At the syllable level, it achieved an overall mean performance (F-score, see section 5.1) of 0.74 (i.e., 74 % mean performance in correct event detection and correct non-event rejection) and 0.53, respectively for boundary detection, and unit detection. In the first task, it is followed by the envelope velocity-based algorithm which had a mean performance of 0.71. Whereas in the unit detection task, it is followed by the simple envelope model which had a mean performance of 0.43. At the word level, the oscillator model achieved an overall mean performance of 0.63 and 0.45, respectively in boundary detection and unit detection. Here again, in the first task, it is followed by the velocity-based algorithm which had the same overall mean performance, and by the envelope model in the second task with a mean overall performance of 0.37.

Unsurprisingly, all models perform better for syllables than for words, and we observe the same pattern in model performance for both syllable and word temporal extraction measures. This could be explained by the fact that all these models are specialized in detecting local extremums of the speech signal envelope, which correspond much more to syllabic boundary markers, whereas this same

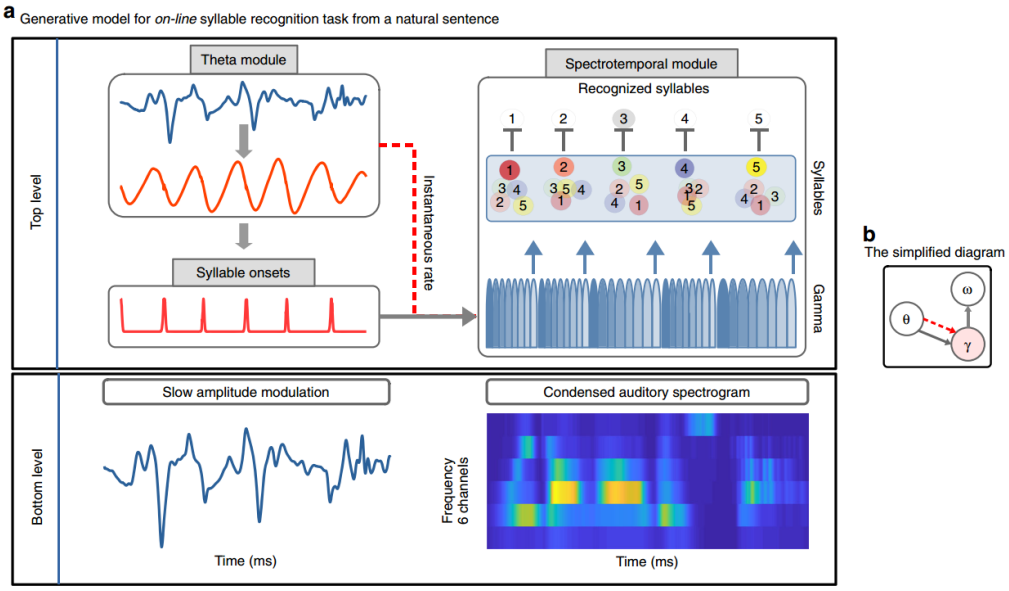

Figure 1.8: Illustration of ***Precoss***, a neuro-computational model of speech perception developed by Hovsepyan et al. (2020). Image is taken from (Hovsepyan et al., 2020).

notion of boundary markers is difficult to define from the signal envelope alone for words.

5.5 *Precoss*: The model by Hovsepyan, Olasagasti & Giraud (2020)

5.5.1 Model description

Figure 1.8 shows the model ***Precoss*** (for "predictive coding and oscillations for speech") developed by Hovsepyan et al. (2020). It illustrates both the full architecture of the model (left part, a panel) and the simplified version of the same model (right part, b panel), depicting only the functional connections from the top level, with the theta module (denoted θ) connected to the gamma module (denoted γ), which in turn connects to the syllable units (denoted ω). This model is conceived as performing the connection between two components described previously: the model of syllabic speech parsing with theta-based envelope modulation processing developed by Hyafil, Fontolan, et al. (2015) and the syllabic decoding model exploiting a Bayesian inference network within dynamic hierarchical generative models developed by Yildiz et al. (2013).

Starting from the raw speech signal, the bottom level extracts the essential characteristics by applying some pre-processings to obtain, on the one hand, the acoustic envelope of the signal (Hyafil & Cernak, 2015; Hyafil, Fontolan, et al., 2015) and, on the other hand, its spectral content (Chi et al., 2005) with

a simplified spectrogram. At the top level, we have two specialized modules mentioned previously, that are the temporal processing model "à la Hyafil", called the theta module, which is stimulated by the slow amplitude modulation of the signal, and the spectro-temporal content processing model "à la Yildiz", aptly named the spectro-temporal module, which is stimulated by the spectral content of the speech input. Taken together, they both contribute to processing the *When* and *What* aspects of speech perception (Arnal, 2012; Arnal & Giraud, 2012).

More precisely, the theta module provides a simplification of the spiking theta model of Hyafil, Fontolan, et al. (2015) by exploiting a continuous non-linear oscillator, the Ermentrout-Kopell's canonical model of Ermentrout and Kopell (1986). It is driven by the signal envelope and resonates in a range of theta frequencies (3–8 Hz). This model tracks so-called theta triggers providing syllable onsets (θ in Figure 1.8, b). From there on, in the spectro-temporal module, each parsed syllable (from 4 to 25 in a single sentence in the corresponding corpus) is fed inside a Bayesian inference network computing prediction error for the corresponding acoustic syllabic signal within dynamic hierarchical generative models respectively testing the match between the incoming signal and each possible syllable in the corpus. These generators perfectly follow the principles proposed by Yildiz et al. (2013) and described previously, with a specific set of 8 states of equal duration per syllable. Therefore, every syllable is encoded by a sequence of 8 spectro-temporal patterns corresponding to 8 gamma units (γ in Figure 1.8, b). At the end of the eighth gamma unit, another sequence of 8 gamma units follows.

Consider now the temporal control mechanisms of the ***Precoss*** model. There are two different mechanisms controlling the sequential activation of syllable units (ω in Figure 1.8, b). The first is through the theta module which provides the syllable onsets detected by the neural tracking mechanisms of the speech envelope by the cortical oscillations. This automatically resets accordingly the 8 gamma units, regardless of the current processing state. The consequence is that, if a syllable onset is detected before the last gamma unit is terminated, then gamma units are nonetheless reset to allow the processing of a new syllable. The second mechanism is through the natural sequence of 8 gamma units, lasting 25 ms each. This sequence can be halted or not, depending on the model detecting another onset before the end of the last gamma unit.

In order to effectively study the model's performance, the authors designed different model variants, based on the pattern of interactions between θ, γ, and ω units. The complete model (model variant A) incorporates the complete processing chain described previously. Then, in a number of variants from B to F, the authors introduce various simplifications to selectively assess the role of each component in the model. First, they degrade the theta module by replacing

its stimulus-driven oscillator with a pure stimulus-driven or a pure oscillating process. Second, they degrade the model by removing one or the other of the temporal control mechanisms for syllable onset detection described in the previous paragraph (one based on direct event detection and the other on the completion of the 8-gamma-unit temporal sequence). Third, they test a model with only fixed sequences of 8 gamma units with no temporal control at all.

5.5.2 Main results

The different model variants were tested on the TIMIT data set (Garofolo, 1993) for syllable recognition using the temporal overlap recognition metric. The TIMIT data set contains phonetically-rich sentences read out by 630 speakers of 8 dialects of American English. The authors used a subset of the initial TIMIT data set amounting to 220 sentences, from 22 speakers.

They first compared all models, except model variant A, that is, the complete model. Results were significantly in favor of model variants with gamma units interacting with omega units (when the evidence accumulation is reset after every full gamma cycle). Within the model variants with the coupling between gamma units and syllable units, model variants with the theta-gamma coupling performed better displaying the importance of having the gamma units' activity synchronized at a preferred rate, whether it be endogenous (200 ms) or exogenous (stimulus-driven theta rhythm). Importantly, slow signal-driven evolution of the mean theta rhythm around 200 ms did not significantly change performance. One can argue that the internal intrinsic theta rhythm (200 ms) captures a wide range of natural syllable rhythms, overall.

In ecological human communication, the speech rate commonly varies from one speaker to another. Nevertheless, listeners usually adapt well to speech rate variability. To study this, the authors performed a simulation where they compressed the original speech input by factors 2 and 3. They compared the whole model A with variant B where the mean theta rhythm slowly adapts to fluctuations of the signal envelope, although with no instantaneous onset detection from the theta module. Interestingly, for a compression factor of 2, there is no significant difference between the two model variants, but for a compression factor of 3, model variant A performs slightly better than model variant B. This suggests that in highly adverse conditions, resetting the sequence of gamma units by a theta module driven by the stimulus envelope is useful, while it is less so in a less adverse situation, where the intrinsic theta rhythm suffices.

6 A key experimental paradigm

After reviewing the main computational models concerned with oscillatory-based processing of speech, we now consider how well they account for human speech perception experimental data. A number of experimental paradigms, mostly involving speech compression, have already been introduced in the previous sections. Here, we focus on key experimental data from Aubanel and Schwartz (2020) that will serve as the basis for our model development and evaluation. Indeed, this paradigm assesses the relative roles of bottom-up resonance processes and top-down linguistic predictions in the onset detection and syllabic parsing mechanism.

6.1 On the role of isochrony in speech perception

Although caveats have been raised about the true periodic nature of cortical oscillations (Cummins, 2012a, 2012b, 2015; Obleser et al., 2017; Obleser et al., 2012), in principle, one would expect that human speech perception, based on these oscillatory cortical processes, would perform best when the speech signal is as regular as possible, i.e., for an isochronous speech signal with temporal events located at regular intervals (Schön & Tillmann, 2015; Tillmann & Lebrun-Guillaud, 2006). Aubanel and Schwartz (2020) tested this hypothesis by conducting an experiment in which they evaluated the intelligibility (word error rate, WER) of the natural speech made isochronous.

6.2 Materials

In their study, the authors used sentences from the *Harvard* corpus for English (Rothauser, 1969), and its equivalent, the *Fharvard* corpus for French (Aubanel et al., 2020) [1]. English and French are two languages with different rhythmic structures, according to the rhythmic class hypothesis (Abercrombie, 1967; Abercrombie, 1965; Grabe & Low, 2002): English is a stressed-timed language and French is a syllable-timed language. Notice that several studies have shown that languages of the world are better classified on a spectrum in this regard, rather than on two discrete categories (Dasher & Bolinger, 1982; Dauer, 1983; Nespor, 1990; Ramus et al., 1999).

Both corpora consist of phonemically-balanced natural spoken sentences uttered by a female speaker for English and by a male speaker for French. Originally, in the French corpus (respectively English) there are 700 sentences, containing 5 (respectively 7) keywords, which are monosyllabic (respectively bisyllabic). However, in their study, Aubanel and Schwartz (2020) randomly

[1] Found online at https://zenodo.org/record/1462854#.YitevozMLm4

selected a subset of 180 sentences per corpora. From these original sentences, they devised three different conditions.

1. The natural rhythm condition, noted NAT, corresponds to unmodified, original sentences of the corpora. Their onset events later serve as a reference for the natural rhythm for modified conditions, either at the accent (acc) or syllable level (syl).

2. The isochronous rhythm condition, noted ISO, corresponds to sentences in which original time onsets are aligned regularly using temporal distortions (detailed below) either at the accent rhythm level or at the syllable rhythm level, using P-centers as units of isochrony in the speech input (Morton et al., 1976; Strauß & Schwartz, 2017). Therefore, there are two variants of this condition: the isochronous condition at the accent level (noted ISO.acc) and the isochronous condition at the syllable level (noted ISO.syl).

3. The anisochronous rhythm condition noted ANI, defined by first reversing the time onsets in the original sentences, and then applying the same isochronous temporal distortions procedure. As for the ISO conditions, there are here also two versions: ANI.acc and ANI.syl. The interest of the ANI conditions is that an "equal" amount of temporal distortion from the natural rhythm was applied as for the corresponding ISO conditions, though not rendering the material isochronous or even temporally more regular than the natural baseline.

Overall, for every sentence, there are 5 different versions in each corpus, corresponding to the different rhythm conditions. The final experimental stimuli consisted of these temporally modified or unmodified sentences, which were also degraded with white background noise at a signal-to-noise (SNR) ratio of -3 dB.

Temporal distortion metric To characterize the departure from isochrony in speech signals, authors used a distortion metric noted δ, as introduced by Aubanel et al. (2016). It is computed for a given reference time series t (the initial temporal event series) which is transformed into a target time series t' (here a hypothetical isochronous time series with the same number of events) as the following:

$$\delta = \sqrt{\frac{\sum_{i=1}^{N} (\log \tau_i)^2 d_i}{\sum_{i=1}^{N} d_i}} \ ,$$

with d_i and d'_i respectively the duration between successive events in the reference and target times series ($d_i = t_{i+1} - t_i$; $d'_i = t'_{i+1} - t'_i$), and τ_i the time-scale factor between the reference and target time series: $\tau_i = d'_i / d_i$.

δ measures the temporal distortion between two time series. If one of the two time series has temporal events that follow a natural temporal distribution,

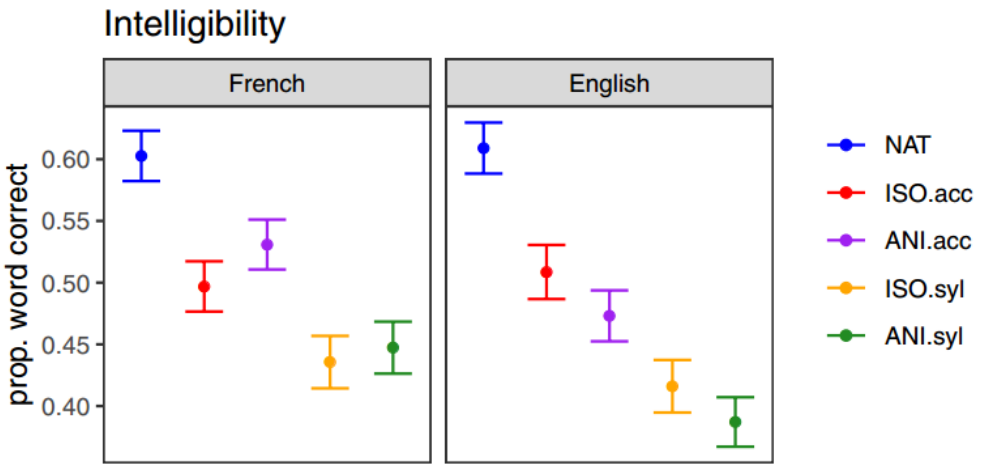

Figure 1.9: Intelligibility results of the Aubanel and Schwartz (2020) experiment in all conditions, for both French and English. Image taken from (Aubanel & Schwartz, 2020).

then δ measures the degradation with respect to naturalness. In that case, the lower δ is, the more natural the sequence of events, and the higher δ is, the more unnaturally temporally distributed the events of the sequence are. If one of the time series is an isochronous time series, then δ measures the degradation with respect to isochrony. In that case, the lower δ is, the more a sequence of events is isochronous, and the higher δ is, the more temporally anisochronous the sequence of events is.

6.3 Results

Subjects' performance was measured as the average number of keywords successfully perceived within the experimental modified or unmodified sentences within the noise.

Figure 1.9 shows a summary of results on intelligibility in all the different conditions, for both languages, French and English. We observe that, for both languages, the naturally unmodified sentences lead to higher correct keyword perception than in any other temporally distorted conditions. Other differences between conditions are mainly attributed by the authors to differences in degradation mechanisms relative to the natural conditions, both in French and English.

To further understand the role of temporal distortion, Figure 1.10 relates the results of intelligibility in the different conditions to various distortion metrics. The main findings are indicated in the annotated regions A, B, and C. First, results in region A show that in the case of unmodified, naturally timed sentences, performance is positively correlated with departure from accent isochrony, whereas

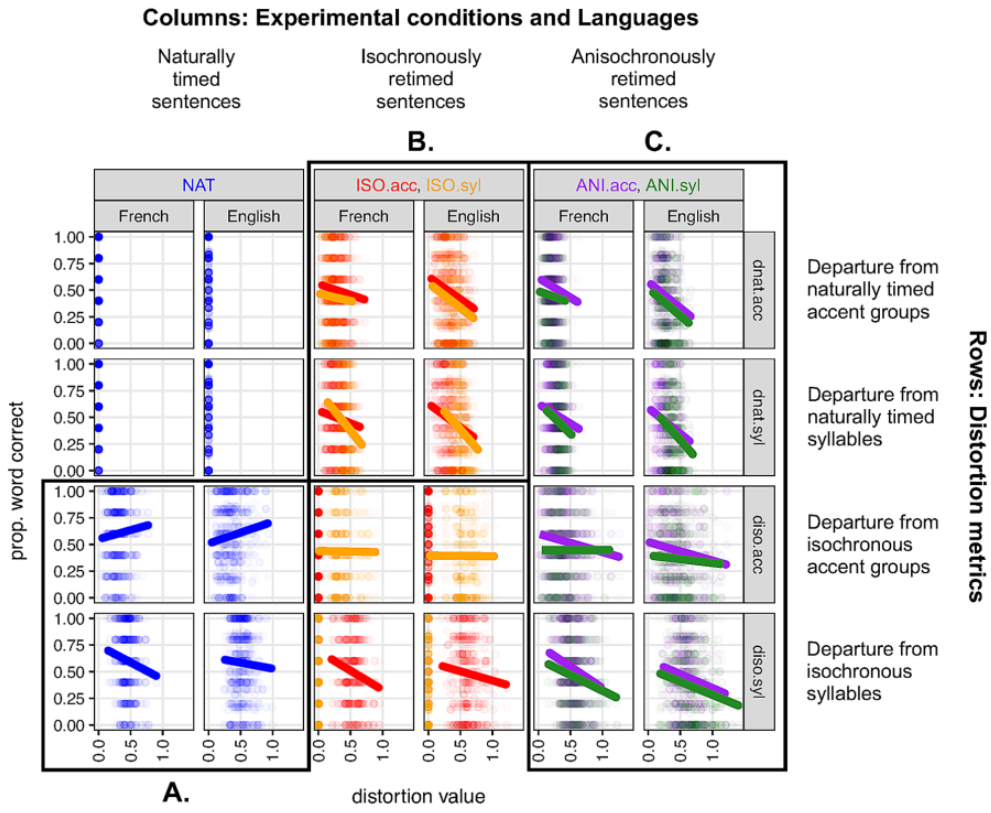

Figure 1.10: Intelligibility results of the Aubanel and Schwartz (2020) experiment in all conditions (columns), in both French and English (subpanels), as a function of various distortion metrics δ (rows). Image taken from (Aubanel & Schwartz, 2020).

it is negatively correlated with departure from syllable isochrony, with the same pattern for both languages, French and English. Importantly, this means that at the syllable level, the more isochronous natural speech is, the more it is intelligible. Second, results in region B show that in the case of isochronously retimed sentences, performance is negatively correlated with departure from naturalness, for both languages, French and English. Finally, region C confirms that for asynchronously retimed sentences, a departure from both naturalness and syllabic isochrony impairs intelligibility.

In conclusion, these experimental data provide two major insights that are (1) that the more speech is naturally timed, the more intelligible it is, and (2) that syllabic isochrony increases intelligibility in naturally time sentences. Importantly, this is true for both languages, notably for English which is not a syllable-timed language. Departure from naturalness in timing appears to result in an overall degradation of intelligibility for the listener.

This could be interpreted by assuming a degradation of the relevance of top-down information from higher-level knowledge, which would allow to better

track the temporal statistics of events in incoming speech embedded in noise. The benefit provided by syllabic isochrony in natural speech could be interpreted as resulting from the oscillatory mechanisms involved in speech perception through cortical oscillations. The interplay between both processes, bottom-up and top-down, as shown by Aubanel and Schwartz (2020), would lead to better intelligibility in speech perception in noise. Altogether, the data and analyses presented by Aubanel and Schwartz (2020) show the interplay between bottom-up and top-down processes in the setting of intelligibility of speech sentences embedded in noise.

7 Interaction between bottom-up envelope processing and top-down predictions in the temporal control of the speech perception process

In this literature review, we have described the two most influential conceptual neuro-computational models of speech perception together with some recent computational models that are compatible implementations. They all agree on some general theoretical principles related to the speech signal processing chain, with an emphasis on the hierarchical oscillatory processing of speech, and a focus on the temporal control of speech processing. The precise mechanism operating such temporal control varies from one model to the other, with a linear oscillator for Räsänen et al. (2018), a nonlinear oscillator for Hovsepyan et al. (2020), excitation-inhibition loops in populations of spiking neurons for Hyafil, Fontolan, et al. (2015) or dynamic sequences of local temporal states for Yildiz et al. (2013). However, crucially, all these models incorporate a specific temporal control mechanism delineating the sequence of spectro-temporal local features to analyze for the decoding of the corresponding speech units. These units are syllables for the first three models, and words for Yildiz et al. (2013). Moreover, the first three models closely relate to the general neural architecture presented by Ghitza (2011) and Giraud and Poeppel (2012) around the theta-gamma coupling associating syllabic parsing in theta units and phonetic decoding in gamma units.

All these models are compatible with the predictive coding framework (Friston, 2005; Friston & Kiebel, 2009; Gilbert & Sigman, 2007; Mumford, 1992; Rao & Ballard, 1999) according to which the human brain is a permanent forward inference machine. The predictive coding framework proposes that sensory, bottom-up processing is compared with top-down, prior knowledge-based predictions. However, in all the computational models that we have described, none precisely integrates top-down knowledge in the temporal control model. For Räsänen et al. (2018), the question does not even arise, because the authors consider the case of children who do not yet have developed lexical knowledge, and hence they focus

on bottom-up mechanisms. For Hyafil, Fontolan, et al. (2015), there is no explicit mechanism of top-down control. All the oscillatory processes implemented are driven by the speech signal itself. For the ***Precoss*** model (Hovsepyan et al., 2020), the situation is slightly different. Clearly, speech segmentation in this model is mainly driven by bottom-up processes with the slow amplitude modulation of the speech input driving the theta oscillator. However, the authors make special use of a parameter called model precision, already introduced by Yildiz et al. (2013) in their model at the basis of spectro-temporal decoding in ***Precoss***. Precision controls the respective weight of top-down evidence and bottom-up incoming features, but it is nevertheless not related to higher-level linguistic (e.g., prosodic, syntactic, or semantic/pragmatic) information. Recently, the authors proposed a new version of ***Precoss*** named "Precoss-β" (Hovsepyan et al., 2022), with the β referring to the potential role of beta oscillations in providing top-down predictions (Pefkou et al., 2017). They suggest that the beta band is involved in controlling the precision of internal states in the decoding process and show that adding this mechanism significantly improves performance over the previous version of the model. It is important to note here that, although this mechanism is apparently related to a top-down mechanism, it does not involve any higher level of stored or learned knowledge. More precisely and of more concern, it is not related to any lexical knowledge.

Generally, whether it is in the field of vision (Lee, 2002; Przybyszewski, 1998), memory (Edin et al., 2009; Gazzaley & Nobre, 2012), motor skills (Narayanan & Laubach, 2006; Roberts et al., 2014), or language (M. H. Davis & Johnsrude, 2007; Perrone-Bertolotti et al., 2012; Radach et al., 2008; Sohoglu et al., 2012), top-down processes actively participate in perception. It is of course also the case for speech perception. To better recognize speech sounds, top-down knowledge (higher-level language systems) associated with morphology, lexicon, syntax, semantics or pragmatics may interact with low-level speech perception processes. Thus, phoneme identification may require a prior understanding of higher lexical units such as words, which might result in an interactive feedback-feedforward process that has been the focus of active research since the first modeling works on spoken word recognition (see e.g., McClelland & Rumelhart, 1981). A classical example concerns the "Ganong effect" (Ganong, 1980; Massaro & Cohen, 1983), in which the identification of an ambiguous stimulus is highly influenced by context. For example, an ambiguous sound that might be a /g/ or a /k/ is more likely to be perceived as a /g/ if followed by "ift" and as a /k/ if followed by "iss". This particular experience has been interpreted as resulting from lexical influence on sub-lexical units (word knowledge influences phoneme perception). More globally, a large number of studies highlight the interaction of top-down knowledge with bottom-up processes, whatever the situation, that is, in adverse (e.g., in noise

Mishra & Lutman, 2014; Zekveld et al., 2006) or normal conditions (Cope et al., 2017; McClelland & Elman, 1986). The "canonical" speech perception model TRACE (McClelland & Elman, 1986) typically integrates the interaction between feedforward and feedback processes, and has been shown in consequence to be able to account for a large set of experimental speech perception data in psycholinguistics, such as the "Ganong effect" discussed previously.

Crucially, however, the nature of top-down processes involved in such models is solely concerned with the "what" question in speech perception, referring to the lexical and sub-lexical unit categorization processes. Still, according to the general neurocognitive framework abundantly developed in the previous sections of this chapter, information regarding "when" and "how" in time the acoustic input is structured and processed is considered encoded independently from "what" and conveyed via distinct neural pathways for both perceptual and motor processes (Arnal, 2012; Arnal & Giraud, 2012; Morillon et al., 2016). In other words, temporal control likely requires not only bottom-up processing of the signal envelope but also top-down predictions from linguistic knowledge. This seems required for at least three sets of reasons. First, if top-down predictions intervene globally in the speech processing architecture, they likely participate not only in the spectro-temporal decoding process but also in the temporal control mechanism driving this process. Secondly, several neural data suggest that the synchronization of brain responses to speech signals does depend on the intelligibility of these acoustic inputs (e.g., Ahissar et al., 2001; Peelle et al., 2013), and it has been repeatedly suggested that top-down predictions could exploit the neural beta channel for this feedforward-feedback process (Hovsepyan et al., 2022; Pefkou et al., 2017). Last but not least, the behavioral data by Aubanel and Schwartz (2020) described in Section 6 seem to clearly show that predictions associated with a "natural" sequence of speech events are required for efficient perception in noise.

8 Goals and contributions of the present thesis

In light of the converging arguments on the potential role of top-down mechanisms in temporal control for speech perception presented in the previous section, the present thesis is focused on three questions related to three main contributions.

The first major contribution of this thesis is the definition and development of a speech perception model, which we named **COSMO-Onset**, that includes both a spectro-temporal content decoding module and a temporal control module **combining bottom-up onset detection and top-down temporal prediction mechanisms**. The separation into two modules is what on the one hand distinguishes our model from classical models of speech perception "à la

TRACE", and on the other hand also allows us to compare with recent models of speech perception inspired by neural oscillations. The specific originality of **COSMO-Onset** compared to the latter models is that it features a temporal control model that combines bottom-up and top-down components. We have defined the model and illustrated its behavior and main principles on a set of simplified stimuli: this contribution was published, first, in a peer-reviewed journal paper to introduce the COSMO-Onset model (Nabé et al., 2021), and second in a peer-reviewed international conference paper to describe the Bayesian Gates, one of its technical features (Nabé, Schwartz, et al., 2022).

Our second contribution in this thesis focuses on the study of a completely bottom-up and simple neural oscillation-based model of speech perception developed by Räsänen et al. (2018). It is a model solely concerned with the detection of syllabic onsets without any consideration related to the spectro-temporal decoding mechanisms. This allows us to show not only the interest and performance in the noise of such systems but also the potential role of resonance and isochrony processes for the robustness of syllabic onset detection in noise. We have also developed and evaluated a variant of the model of Räsänen et al. (2018), that detects P-centers instead of syllabic onsets. This second contribution was published in a peer-reviewed conference (Nabé, Diard, et al., 2022).

The third and last contribution of this thesis is the development of the second version of **COSMO-Onset** integrating more advanced signal processing and decoding mechanisms, to be able to deal with real speech signals. To do so, some simplifying assumptions of the first version of **COSMO-Onset** have to be lifted, and some of its components need to be redesigned. For instance, since the model developed by Räsänen et al. (2018) already deals with real speech input, it can be adapted and included into the **COSMO-Onset** as its bottom-up onset detection mechanism, in the temporal control module. With this version of **COSMO-Onset** adapted to real speech, we then perform a first complete simulation of the experiment of Aubanel and Schwartz (2020) and evaluate whether **COSMO-Onset** successfully accounts for the complementary roles of isochrony and naturalness in speech perception in noise.

Chapter 2

COSMO-Onset: The conceptual model

Note

This chapter is partially adapted from (Nabé et al., 2021).

In the previous chapter, we conducted a literature review that provides an overview of the research on neural oscillations and their specific relations to speech. In particular, we described the main models of speech perception and segmentation based on neural oscillations. This chapter ended with a critical analysis of these models which allowed us to define the main theoretical questions of this thesis.

The main contribution of this thesis is the COSMO-Onset model. We have developed two implementations of this model. The first one, presented Chapter 3, served as a proof-of-concept of the proposed architecture and inference mechanisms and was designed to process very simplified, synthetic stimuli. The second one aimed at processing more realistic stimuli. Its description is covered in Chapters Chapter 4, Chapter 5 and Chapter 6. Therefore, in the present chapter, we introduce COSMO-Onset, focusing on a conceptual level overview (i.e., with no mathematical details) and elements that are common to the subsequent implementations.

First, we present the overall model architecture and its two main components: the decoding module and the temporal control module. The decoding module processes the spectro-temporal content of the input hierarchically in order to ultimately recognize words. The temporal control module processes the slow amplitude modulation of the input to detect syllabic events, by combining both sensory-driven bottom-up segmentation and knowledge-driven top-down event prediction.

Second, we describe, again at a conceptual level, the Bayesian inference process conducted in the model for simulating word recognition and show how it articulates with and involves syllabic event segmentation. We detail the probabilistic computations featured in the temporal control and decoding modules.

1 Model architecture

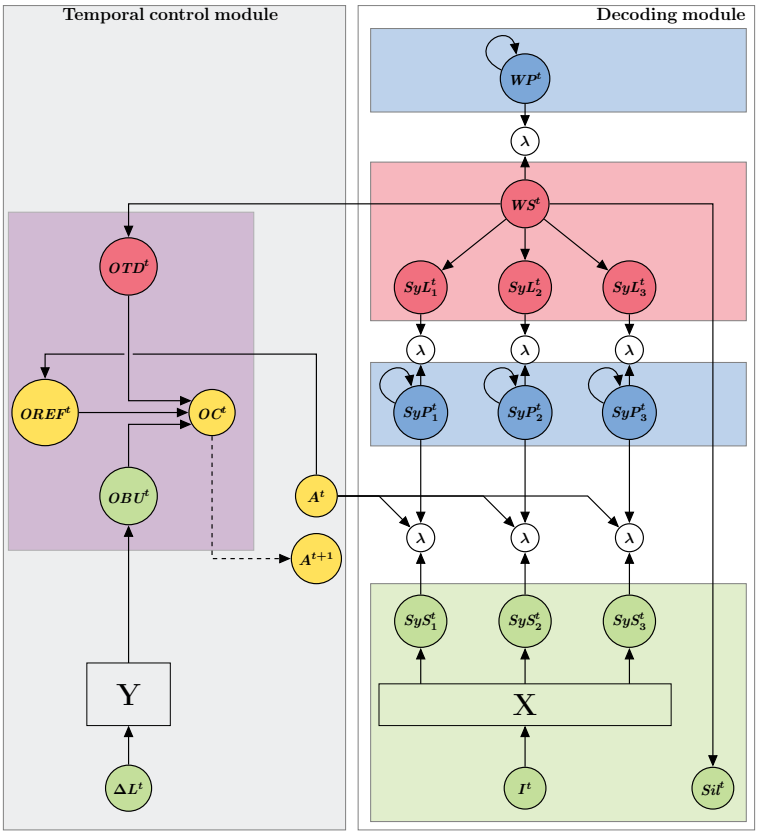

Figure 2.1: Conceptual graphical representation of COSMO-Onset

COSMO-Onset is a Bayesian hierarchical computational model of speech perception. The conceptual architecture of the model, which is graphically represented on Figure 2.1, shows the main components: on the left is the "temporal control" module, and on the right is the "decoding" module. The temporal control module associates three components. A bottom-up system associates input envelop features based on variations of the perceptual intensity (loudness) ΔL with Onset Bottom-Up OBU features. A top-down system predicts the Onset of Top-Down OTD features from the output of the word recognition module. Finally, the bottom-up and top-down systems are combined (purple box in Figure 2.1). The decoding module hierarchically combines a sensory layer

decoding sensory inputs I into "Sensory" Syllabic features SyS (green box in Figure 2.1) and a perceptual layer associating "Perceptual" Syllabic features SyP into Word units W (blue box in Figure 2.1).

A combination tool used in the model to connect some of its portions is the so-called coherence variables (noted as λ variables), which will be further described in the next sub-sections. The summary of the different variable names and their interpretation is provided in Table 2.1. Applying Bayesian inference to the model architecture provides computing steps for simulating onset detection and word recognition. We now describe the general principles and functioning of the model followed by a description of each of the two sub-models.

Table 2.1: Summary of symbols of the illustrative COSMO-Onset model: variable names and their interpretation. Some of the mentioned variables (e.g., *Sil*, *FeP*, ...) will be discussed later in Chapter 3

Variables for acoustic signal description	
$I_{1:12}^{1:T}$	Spectro-temporal content of the acoustic signal Input (F1, F2 formants)
$\Delta L_{1:12}^{1:T}$	Derivative of the Loudness local intensity of the acoustic signal
$Sil_{1:12}^{1:T}$	SILent portions of the acoustic signal (Boolean)
Variables for linguistic content	
$FeS_{1:12}^{0:T}$, $FeP_{1:12}^{0:T}$, $FeL_{1:12}^{0:T}$	Phones (i.e., FEatures), respectively from Sensory decoding, in phone Perceptual accumulators and from Lexical prediction
$SyS_{1:3}^{0:T}$, $SyP_{1:3}^{0:T}$, $SyL_{1:3}^{0:T}$	SYllables, respectively from Sensory decoding, in syllabic Perceptual accumulators and from Lexical prediction
$WS^{0:T}$, $WP^{0:T}$	Words, respectively from Sensory decoding and in the word Perceptual accumulator
Variables for controlling information flow in the model	
$\lambda FeSP_{1:12}^{1:T}$, $\lambda FePL_{1:12}^{1:T}$, ...	Coherence or controlled coherence variables, connecting layers of the model
$A_{1:15}^{1:T}$	Control variables modulating information flow (opening, closing, and sequencing phone and syllable perceptual accumulators)
Variables for onset detection	
$OTD^{1:T}$, $OBU^{1:T}$, $OREF^{1:T}$, $OC^{1:T}$	Onset detectors (Boolean), respectively from TD knowledge, BU sensory decoding, REFractory period inhibition, and Combined result

1.1 General principles: Coherence variables, Bayesian gates, and syllabic parsing

As shown in Figure 2.1, the overall structure of the decoding portion of the model consists of different layers, connected by Boolean variables called "coherence variables" (represented by λ nodes in Figure 2.1). These can be seen as "probabilistic glue", allowing merging, in a mathematically principled manner, probability distributions over the same domains (Bessière et al., 2013; Gilet et al., 2011). During inference, these coherence variables are used to choose how probabilistic information propagates into the model; in that sense, they can be interpreted as "Bayesian switches" that can be "closed" or "open". When a coherence variable is closed between two connected variables, information propagates through it. Mathematically, this corresponds to assuming that the value of the coherence variable is known and equal to 1, and it yields a product of the probability distributions of the variables connected by the coherence variable (whatever these probability distributions). Conversely, a Bayesian switch can be "open", by ignoring its value during inference; this results in disconnecting the corresponding portions of the model connected by the coherence variable, through a marginalization process that can be shown to simplify. Technical details can be found elsewhere (Gilet et al., 2011, see also Section 4).

Some of the coherence variables in the decoding module (the ones with input arrows coming from node A^t in the temporal control module, see Figure 2.1) are further "controlled" (Phénix, 2018), that is to say, they allow controlling in a gradual manner the propagation of probabilistic information, from one layer to another. Where coherence variables can be interpreted as "Bayesian switches", controlling information flow in an all-or-none manner, controlled coherence variables can be interpreted as "Bayesian potentiometers", thanks to their gradual control of information propagation. Technically, this is done by connecting a probability distribution over the control variable, which is Boolean, to the coherence variable. The probability that the control variable is "*True*" then modulates the amount of probabilistic information propagation (see Section 4).

In the context of the COSMO-Onset model, we, therefore, use controlled coherence variables to modulate, over time, information flow in the model (see variable A^t in Figure 2.1). This allows us to modify dynamically, during perception, which portion of the model receives and processes sensory evidence. In other words, variables A^t, which are the main output of the temporal control module, are used to explicitly "open" or "close" channels through which probabilistic information propagates in the model.

More precisely, we employ such a mechanism to control information flow between the acoustic input and a number of consecutive "syllabic decoders". In the following, we fix this number to 3 for simplicity, which means that the word

decoder includes a sequence of 3 syllabic decoders (associated with the variables SyP_1^t, SyP_2^t and SyP_3^t). In this process, control variables A^t control the temporal windows during which syllable perceptual variables receive sensory evidence to process, from the acoustic input. This allows implementation of the sequential activation of syllabic decoders, that is to say, syllabic parsing.

To generalize, we propose to call "Bayesian gates" this novel mathematical construct (Nabé, Schwartz, et al., 2022). They allow the segmentation of a sensory stream by appropriately activated perceptual decoders in a sequential manner. In short, variables $A_{1...3}^t$ are in charge of controlling which syllable decoder receives sensory information. Their probability distributions pilot all links between decoders and the sensory input. When the probability that the control variable A_i^t is *True* is 0, the corresponding decoder i is not activated or already terminated; on the other hand, when the probability that the control variable A_i^t is *True* has a non-zero value, the corresponding decoder is currently activated, so that some amount of sensory information is fed into the perceptual model.

The purpose of the temporal control module is thus exactly to control syllabic parsing. To do so, it computes, at each time step, the probability that there would be a syllabic onset event, that is, the probability that a new syllable begins in the acoustic input. When this probability passes a threshold, the system decides there was an onset, which has two main effects (see Figure 2.1, dotted arrow). First, the currently "activated" syllable decoder stops receiving sensory input from the stimulus or lower-level layers. Second, the next syllabic decoder, in sequential order, is activated. Therefore, our model segments the continuous speech stream into linguistic intervals of varying lengths at the syllabic level. Consequently, the model can handle words that have a varying number of syllables, and syllables that have a varying duration. We note, as mentioned in the introduction (Section "Neural oscillation-based models of speech segmentation"), that previously proposed models also feature such mechanisms, of sequential activation and deactivation of syllabic decoders (Ghitza, 2011; Hovsepyan et al., 2020; Hyafil, Fontolan, et al., 2015).

1.2 Decoding module

The decoding module of the COSMO-Onset model is inspired both by the BRAID model (Bayesian model of Word Recognition with Attention, Interference, and Dynamics) of visual word recognition (Ginestet et al., 2019; Phénix, 2018) and by the classical, three-layer architecture of spoken word recognition models, such as TRACE (McClelland & Elman, 1986; McClelland & Rumelhart, 1981). It can also be construed as a hierarchical (multi-layered) dynamic Bayesian network (Murphy, 2002), with an external component to control information propagation.

With its hierarchical architecture, each layer describes particular knowledge that is involved in the overall word recognition process. From the first bottom layer to the last top layer, the different layers can interact with each other, either in a purely bottom-up manner with only feedforward processes or in a bi-directional manner with feedforward and feedback processes. The model is organized into different layers of knowledge representation. This begins with a syllabic sensory layer (the green box in Figure 2.1), connected to a syllabic perceptual layer (in blue), which is then connected to a lexical layer (in red) expressing the known word syllabic constitution (word-to-syllable lexical layer), which is finally connected to a word perceptual layer (in blue).

The **sensory layer** is represented by the rectangle noted **X** Figure 2.1, and contains a mathematical relation between syllables and sensory input; more precisely, it represents knowledge about how known syllables correspond to acoustic signals. This can be expressed in a general mathematical form by a probability distribution of the form $P(SyS_i^t \mid I^t)$, that is, "what is the probability distribution for the *ith* syllable in the sensory layer at time t, knowing the input I^t at the same time?".

There are two **perceptual layers** featured in the COSMO-Onset model, with internal representations corresponding to syllables and words. Each is associated with a series of probabilistic dynamic models (i.e., Markov-chain-like probabilistic models, to which we now refer as "decoders"). They are expressed in terms of the forms $P(SyP_i^t \mid SyP_i^{t-1})$ and $P(WP_i^t \mid WP_i^{t-1})$, respectively for syllables and words. They allow continuous accumulation of sensory evidence about the representation domain they consider. Information gradually decays from these Markov chains, to ensure a return to their respective initial states in the absence of stimulation. However, the information decay rate is set to a low value, to basically ensure that the result of sensory evidence accumulation is maintained and remains available for the whole duration of processing a given word. Therefore, these Markov chains essentially provide perceptual models about syllables and words, central to syllable and word recognition, respectively.

In the **lexical layer**, the model contains probabilistic "transformation terms", that is to say, knowledge about how one representational space maps onto another. The word-to-syllable lexical layer describes how known words are composed of known syllables. This is expressed with terms of the form $P(SyL_i^t \mid WS^t)$. With a similar mechanism, known syllables can be described in terms of smaller units such as phones. This provides a syllable-to-phone lexical layer in a similar manner to the word-to-syllable layer. This is not included in the conceptual representation shown on Figure 2.1, though it appears later in Chapter 3.

These different layers are connected by coherence variables so as to have sensory information entering the model through the bottom sensory layer and

propagating to the top last perceptual layer of word recognition. Altogether, we can then simulate word recognition with the decoding portion of the model, that is to say, compute the probability distribution over variable WP^t, at each iteration t, given the acoustic stimulation, as described by variables I^t, Sil^t and ΔL^t. Because of the complex structure of the model, with its hierarchically layered Markov chains, Bayesian inference results in complex computations, involving both feedforward (from acoustic input to word space) and feedback (from word space to acoustic input) propagation of information. However, we approximate these, considering word recognition in the decoding module as a pure feedforward process (in contrast with our main focus, that is, the inference in the temporal control module, which features both bottom-up and top-down components; see below).

1.3 Temporal control module

The temporal control module is composed of three interacting portions, with "bottom-up" onset detection, "top-down" onset prediction, and the last portion to combine them.

The **bottom-up portion** of the temporal control module assesses the probability of syllabic onset events by relying on the temporal cues that can be extracted from the speech envelope. The extraction mechanisms are described by the model in the rectangle noted **Y**. This has been largely discussed in the literature and several models of syllable onset and boundary detection have been proposed (Ghitza, 2011; Hovsepyan et al., 2020; Hyafil, Fontolan, et al., 2015; Mermelstein, 1975; Räsänen et al., 2018). These models process the speech envelope, in search of either rapid increase (towards peaks) or decrease (troughs) in the energy of the speech envelope. In practice, the **Y** rectangle can contain any of the syllable event detection models, be they based on oscillations or other signal processing techniques. This bottom-up process outputs OBU^t events, that can either be precisely or probabilistically positioned in time in the input signal.

The **top-down portion** of the temporal control module relies on lexical knowledge about word composition. Such lexical knowledge associates each word of the lexicon to a sequence of syllables, each of a known composition, thus of typical known duration. Therefore, the model incorporates knowledge about the "canonical" instants at which syllabic onset can be expected, for each word. This is expressed in terms of the form $P(OTD^t \mid WS^t)$. During word recognition, this lexical prediction of onset events is combined with the ongoing computation of the probability distribution over words, so that words contribute to syllabic onset prediction according to their estimated probability. Other levels allowing top-down predictions on syllable duration and event occurrence (e.g., syntactic or prosodic), could be proposed, and will indeed by introduced or discussed

later in this document (in Chapter 6 and Chapter 7). At this stage, for the sake of simplicity, we just feature in the COSMO-Onset model lexical top-down predictions.

The next component of the temporal control module is a **fusion model** between the bottom-up detection and top-down prediction of onset events. It is expressed by the term $P(OC^t \mid OTD^t\, OBU^t)$. We define two ways of combining the two pieces of information, through two fusion "operators", the *AND* and the *OR* operators. They are both mathematically defined as particular products of the probability distributions provided by the top-down and bottom-up components. Nevertheless, they can easily be interpreted: with the *AND* operator, the temporal control module decides that there is an onset event if both the bottom-up OBU^t and top-down OTD^t components agree that there is one; in contrast, with the *OR* operator, the temporal control module decides that there is an onset event if at least one component suggests that there is one.

The computed probability that there is an event, noted by the term $P([OC^t = \text{True}] \mid OTD^t\, OBU^t)$, is then compared with a decision threshold: if it exceeds this threshold, an onset event is considered detected. This then controls, sequentially, the closing and opening of the Bayesian Gates connected to the appropriate syllable decoders of the decoding module. It is represented, in Figure 2.1, by the dotted arrow between nodes OC^t and A_i^{t+1} (with $i = 1 \ldots 3$), implementing the Bayesian Gate process explained in Section 1.1.

2 Inference for simulating word recognition

The main cognitive task we want to simulate with the COSMO-Onset model is word recognition. These computations are performed online (continuously as the model receives input), thanks to the recursive solutions provided by Bayesian inference.

Both syllable event detection and word recognition are computed at each time step: word recognition proceeds assuming the states of the syllable decoders as given, and event detection, informed by word recognition, proceeds to compute the states of syllable decoders for the next time step. In other words, model simulation proceeds in an iterative manner, as only probability distributions at time t are needed to compute probability distributions at time $t + 1$ (notice that even though, for visualization purposes, we also memorize the whole history of probability distributions, this is not required for simulations).

2.1 Inference in the decoding module

Formally, **word decoding** relies on **syllable decoding**. To simulate these, we compute the probability distributions over the perceived syllables SyP and word

WP, at each time step, assuming that the stimulus and states of each syllable decoder (i.e., whether they are active or not) are given. To differentiate these two computations, we use the coherence variables to limit the propagation of information extracted from the stimulus into the model.

Consider first **syllable decoding**. For a syllable decoder i ($= 1 \dots 3$), it is represented by computing a probability distribution $\mathcal{Q}Sy_i^t$, expressed in terms of the probabilistic variables in the sensory and perceptual layers at the syllable level (in all the following, the introduction of the calligraphic letter $\mathcal{Q}$ in a given variable name means that it is the result of a probabilistic question, e.g. "what is the content of a given variable knowing a given state of other available variables?"). The exact computation of $\mathcal{Q}Sy_i^t$ depends on the **X** box, and it is not detailed in this part. The computation is performed over the entire speech stimulus presentation, even if there are some moments (time steps) when the syllable decoder is not active. In those moments, the probability distribution decays towards its initial state of uniform knowledge about all the syllables states.

Next and finally, in a similar manner, for **word decoding**, we compute a particular probability distribution, noted $\mathcal{Q}W_i^t$. It is detailed for each of our implementations of the COSMO-Onset model, in the following chapters.

2.2 Inference in the temporal control module

At each time step, once word decoding is computed, we then compute, in the temporal control module, onset detection, to update the states of syllable decoders for the next, upcoming time step. **Inference for the bottom-up sensory detection of onsets** simply proceeds by referring to the $P(OBU^t \mid \Delta L^t)$ term, with computation mechanisms encoded in the **Y** box of Figure 2.1. On the other hand, for the **inference of the top-down prediction of onsets**, we compute the probability:

$$P([OTD^t = True]) = \sum_{ws^t} P([OTD^t = True] \mid [WS^t = ws^t])\ \mathcal{Q}W^t\ .$$

In other words, we compute the probability that there would be an onset, according to the lexical models of all words, simultaneously, but weighed according to the current probability distribution over words as computed by word recognition. This, consequently, allows word perception to influence syllabic onset detection at any time of the inference.

Finally, we define the **fusion operator**, expressed by the term $P(OC^t \mid OTD^t\ OBU^t)$. In the *AND* variant of the fusion operator, we define:

$$\begin{aligned} &P([OC^t = T] \mid [OTD^t = T]\ [OBU^t = T]) \\ &= \quad P([OTD^t = T]) \times P([OBU^t = T])\ , \end{aligned}$$

with T for the *True* Boolean value. This implements a combination in which the probability of onset is the product of the probabilities of the two "temporal submodels" (top-down prediction and sensory detection). As a consequence of this product, the probability value can be close to one only when the two components agree and also provide a high probability that there is an onset; this explains why we denote this as an "AND" combination operator.

To define the OR operator, we apply De Morgan's law, $A \vee B = \overline{\overline{A} \wedge \overline{B}}$, and define:

$$\begin{aligned} & P([OC^t = T] \mid [OTD^t = T]\ [OBU^t = T]) \\ & \quad = \quad 1 - (1 - P([OTD^t = T])) \times (1 - P([OBU^t = T])) \ . \end{aligned}$$

The final step of onset detection is to apply the decision process on the computed probability distribution: when the probability that $[OC^t = True]$ is above a threshold, an onset is considered to be detected, which updates the states of syllable decoders. This final step is not properly a "probabilistic dependency" in the model; that is why it is represented as a dotted arrow in Figure 2.1.

3 Discussion

In this chapter, we presented a conceptual description of the COSMO-Onset model. We defined an original Bayesian model of speech perception including a temporal control module combining bottom-up acoustic envelope processing and top-down timing predictions from higher linguistic levels.

Even though we did not specify the exact mechanisms at play for syllable decoding and syllable event detection, it led us to lay down the key concepts and design two fusion models associating bottom-up event detection and top-down temporal prediction. We now have at our disposal a general Bayesian architecture associating temporal control and input decoding, that can be implemented, tuned, or modified in various ways, tested on speech stimuli, and possibly used for generating predictions for future neurocognitive experiments.

Chapter 3

COSMO-Onset: The illustrated model

Note

This chapter is partially adapted from (Nabé et al., 2021).

In the previous chapter, we presented the COSMO-Onset model at the conceptual level. We specifically defined all the main components and their interaction, leaving out some details to be specified for each implementation of the model. To recall, we note **X** the portion of the decoding module which maps syllables onto sensory input, and **Y** the portion of the temporal control module which implements the syllable event detection mechanism.

In this chapter, we first describe the two boxes **X** and **Y** for our first implementation of the model. We then perform various simulations in an exploratory study aiming to illustrate the model behavior. These simulations rely on "toy", synthetic stimuli, and experimental situations, that were carefully designed to assess model performance in nominal conditions, and evaluate its robustness when confronted with various adverse conditions.

1 The illustrated COSMO-Onset model

After the conceptual overview of the model in the previous chapter, we now introduce the first implementation of COSMO-Onset, with which the simulations of this chapter will be performed. The probabilistic dependency structure of the model is shown on Figure 3.1.

In this version, in the **X** portion of the **decoding module**, syllables are decomposed into phones, with phones being short portions of the acoustic input (possibly close, but not always equivalent to a phoneme). We assume, for

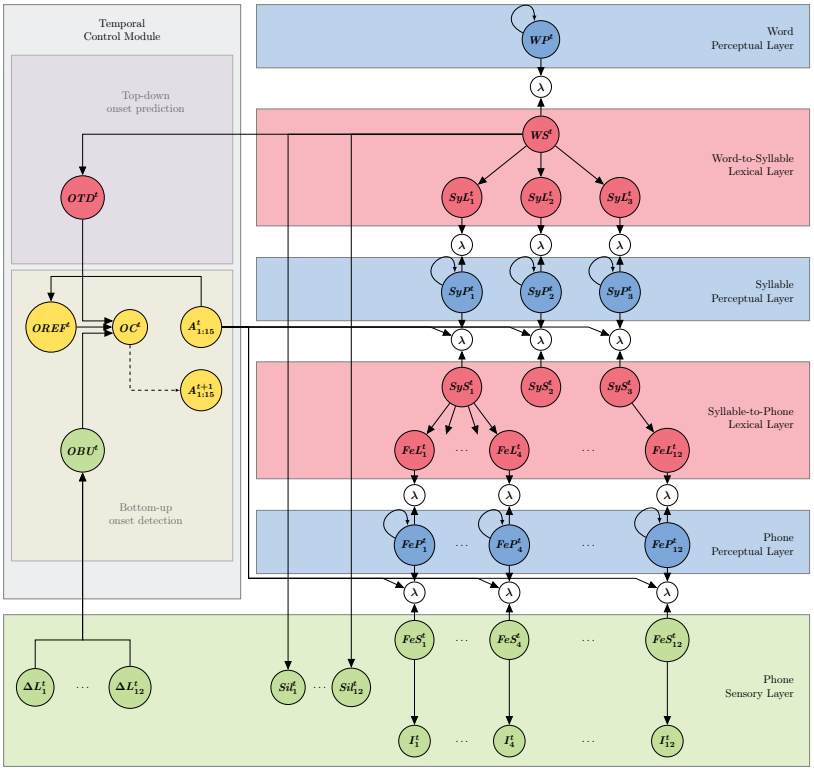

Figure 3.1: Graphical representation of the first implementation of the COSMO-Onset model. Variables of the model are represented as nodes (A summary of variable names and their interpretation is available in Table 2.1). Subscripts indicate the position in sequential parsing of the input into the linguistic unit, and superscripts indicate time instant. For instance, SyP_1^t is the variable related to the first syllabic decoder at time t. Probabilistic dependencies between variables are represented by arrows: there is an arrow from node X to node Y if X is a "parent node" of Y, that is to say, X appears as a conditioning variable of Y in a term (e.g., the arrow from SyS_1^t to FeL_1^t represents the term $P(FeL_1^t \mid SyS_1^t)$). Self-looping arrows denote dynamical models, that is to say, a variable that depends on the same at the previous time step (e.g., there is a term $P(WP^t \mid WP^{t-1})$). The dotted arrow between node OC^t and node $A_{1:15}^{t+1}$ is not a probabilistic dependency and represents instead a decision process (*i.e.*, the probability that OC^t is *True* is compared to a threshold, and this comparison conditions variables $A_{1:15}^{t+1}$). Sub-models are represented as colored rectangular blocks, to assist model description (see text for details). Portions of the model, specifically, some "phone branches" are not shown, for clarity.

simplicity, that there are not more than 4 phones per syllable. In consequence, the decoding module now incorporates a **phone perceptual layer** and a **syllable-to-phone lexical layer**. As for the other two perceptual layers (word and syllables), the **phone perceptual layer** has internal representations corresponding to phones and is associated with a series of probabilistic dynamic models, expressed in terms of the form $P(FeP_j^t \mid FeP_j^{t-1})$. This layer is connected to the **phone sensory layer** with coherence variables.

In the **Y** portion of the **temporal control module**, we assume that there is a straight dependency between the speech input loudness and the bottom-up onset detection variable. In other words, our onset detection mechanism is a direct computation from the loudness profile, adequate in the context of our synthetic stimuli. Hence, there is no external onset detection mechanism in this version of the model, and we apply a computation directly derived from the loudness. The syllable event detection from the stimulus is based on tracking the rapid increase of energy in the speech envelope. If such an increase is detected for several successive time steps, and if the corresponding increase exceeds a given threshold, then the probability of an onset event gets high. This is expressed in terms of the form $P(OBU^t \mid \Delta L^t)$. However, since this simplified bottom-up onset detection method is not oscillatory per see (in the sense of neural oscillatory models of speech perception), we need a mechanism to avoid detecting two successive events too close. Therefore, we introduce a refractory period.

At this stage, the **refractory period** is expressed by the variable $OREF^t$. If a syllable event was detected, this refractory mechanism prevents successive detection for a given time window, fixed at 50 ms. This is inspired by well-known properties of the dynamics of oscillators in speech processing, that prevent the firing of successive onsets in the same oscillation period, classically observed in the theta band (Schroeder & Lakatos, 2009; Wyart et al., 2012). This is also a classical feature of previous models (Hyafil, Fontolan, et al., 2015).

Consequently, according to this description of the temporal control module of the first implementation, we adapt the fusion equations. In the *AND* variant of the fusion operator, we now define:

$$\begin{aligned} &P([OC^t = T] \mid [OTD^t = T]\ [OBU^t = T]) \\ &\quad = \quad P([OTD^t = T]) \times P([OBU^t = T]) \times P([OREF^t = T])\ , \end{aligned}$$

and the *OR* variant is defined as follows:

$$\begin{aligned} &P([OC^t = T] \mid [OTD^t = T]\ [OBU^t = T]) \\ &\quad = \quad (1 - (1 - P([OTD^t = T])) \times (1 - P([OBU^t = T]))) \times P([OREF^t = T])\ . \end{aligned}$$

2 Simulation Material

In this first set of simulations aiming at evaluating the feasibility of the global model and exploring the potential role of top-down temporal predictions, we defined a set of highly simplified materials, easily tractable but still enabling to test all the different components of the model. This "toy" material hence respects a compromise between two antagonist requirements: being sufficiently

varied to display a variety of configurations for the model and being sufficiently simple to enable simple simulations that remain easy to interpret.

2.1 Linguistic material

The linguistic material we consider in this first study is made of isolated words with a variable number of syllables from 1 to 3, and syllables made of either a single vowel (a V syllable) or a sequence of a consonant and a vowel (a CV syllable). We consider a set of 3 vowels /a i u/ and 2 plosive consonants /p t/. Furthermore, we defined a lexicon of 28 toy words, at most tri-syllabic, the list of which is provided in column 2 in Table 3.1.

2.2 Phonetic material

At the acoustic and phonetic level, we represent syllables by sequences of phones with a maximum number of 4 phones per syllable (in the same vein as in Ghitza (2011)). The sequence of phones for the 28 words in the lexicon is provided in column 3 in Table 3.1. For example, the word *"pata"* is composed of a sequence of 7 phones *p-@-a-t-@-a-#*, the content of which will be described in the following of this section. Altogether, the constraints on the maximal number of syllables per word (3) and phones per syllable (4) match with the decoding structure of the current COSMO-Onset implementation (see Figure 3.1), respectively in the word-to-syllable lexical layer and syllable-to-phone lexical layer.

Vowel and plosive phones in our simulations are acoustically represented as sets of pairs of formants *(F1, F2)* in Barks, a subjective perceptual scale (Zwicker, 1961) (see Figure 3.2). While it is classical to characterize vowels by their first two formants (Fant, 1970), it is less classical to use formant values for plosives (although, see Schwartz et al. (2012) for characterization of plosives by formant values). More precisely, the formant values for the considered vowels are gathered from a dataset obtained using VLAM, the Variable Linear Articulatory Model (Boë & Maeda, 1998; Maeda, 1990). It contains a large set of synthetic acoustic samples for all oral French vowels, and we only used the data points for the vowels /a i u/, respectively corresponding to phones *a*, *i* and *u* in the following, which amount to 15,590 samples. To this vowel set, we added 1,000 points for the phones *p* and *t* associated with consonants /p t/, 500 each, supposed to lie in the *(F1, F2)* space between *i* and *u*, *p* close to the back rounded *u* and *t* close to the front *i* (Schwartz et al., 2012).

For the syllables formed by two different phonemes (in the present simulations, C followed by V), in order to simulate formant transitions (Dorman et al., 1975; Lindblom & Studdert-Kennedy, 1967; Stevens & Klatt, 1974), we defined linear transitions between the phones associated with the constituent phonemes.

Table 3.1: List of the 28 words of the lexicon together with their "phonetic" content. Column 1 provides the grouping of words according to their number of syllables. Column 2 provides the name of each word, corresponding to its phonological content. Column 3 provides the phonetic content of each word, that is, the sequence of acoustic phones. Column 4 provides the corresponding duration of the model input, in simulated time steps.

Word type	Word	Phone sequence	Duration
Monosyllabic	*"a"*	*a-#*	150
	"pa"	*p-@-a-#*	200
	"pi"	*p-@-i-#*	200
	"pu"	*p-@-u-#*	200
	"ta"	*t-@-a-#*	200
	"ti"	*t-@-i-#*	200
	"tu"	*t-@-u-#*	200
Bi-syllabic	*"apa"*	*a-p-@-a-#*	300
	"ata"	*a-t-@-a-#*	300
	"ipi"	*i-p-@-i-#*	300
	"iti"	*i-t-@-i-#*	300
	"upu"	*u-p-@-u-#*	300
	"utu"	*u-t-@-u-#*	300
	"papa"	*p-@-a-p-@-a-#*	350
	"pata"	*p-@-a-t-@-a-#*	350
	"patu"	*p-@-a-t-@-u-#*	350
	"pipi"	*p-@-i-p-@-i-#*	350
	"pita"	*p-@-i-t-@-a-#*	350
	"tata"	*t-@-a-t-@-a-#*	350
	"tatu"	*t-@-a-t-@-u-#*	350
	"tuti"	*t-@-u-t-@-i-#*	350
Tri-syllabic	*"apata"*	*a-p-@-a-t-@-a-#*	450
	"apiti"	*a-p-@-i-t-@-i-#*	450
	"iputu"	*i-p-@-u-t-@-u-#*	450
	"utatu"	*u-t-@-a-t-@-u-#*	450
	"patata"	*p-@-a-t-@-a-t-@-a-#*	500
	"patati"	*p-@-a-t-@-a-t-@-i-#*	500
	"tapatu"	*t-@-a-p-@-a-t-@-u-#*	500

Examples of transitions between phones p and a and between phones a and t are depicted in Figure 3.2. Transitions are denoted by the phone symbol @, both in descriptions of the stimuli used in the simulations, but also as a value in the phone space in the model.

Finally, in the present simulations, each word input consists of a phone sequence ending with an "end of sequence" marker, to signal silence in the acoustic signal. Silence is denoted by the phone symbol #, here again, both in descriptions of the stimuli and as a possible phone to be recognized by the model. An example of formant sequence used as input for the bi-syllabic word "*pata*" is

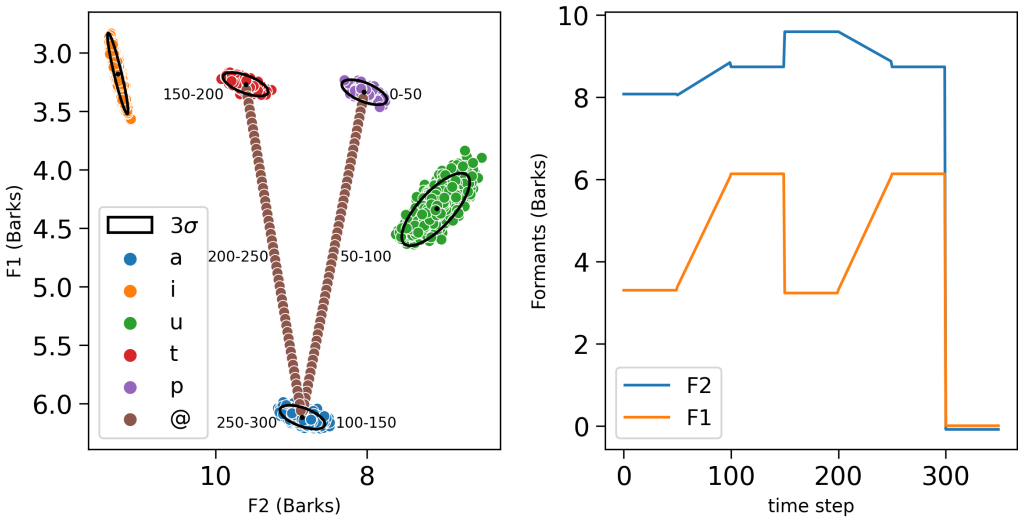

Figure 3.2: (Left) Phones of the lexicon represented on a two-dimensional space with the second formant F2 on the x-axis from right to left and the first formant F1 on the y-axis from top to bottom, as is classical in phonetic displays. Phones associated with phonemes /a/, /i/, /u/, /t/, and /p/ are respectively represented by blue, yellow, green, red, and purple colored dots. The trajectory of the simulation of the word "pata", is also displayed. The annotations correspond to the different corresponding time steps for each constituent phone of the word "pata", with one sample of the phone *p* from 0 to 50 time steps, the transitional phone @ between the phones *p* and *a*, from 50 to 100 time steps, along a linear transition joining the barycenters of the two phone categories (brown dots), one sample of the phone *a* from 100 to 150 time steps, one sample of the phone *t* from 150 to 200 time steps, the transitional phone on 50 time steps, and one sample of the phone *a* from 250 to 300 time steps. For each phone, are also shown the mean (black dot) and the 3 standard-deviation ellipses of the bi-variate normal distribution best fitting the data points (black ellipses). (Right) Example of formant inputs (*y*-axis, in orange for F1, in blue for F2) used for the word "*pata*", as a function of simulated time steps (*x*-axis).

shown in Figure 3.2 (right).

All these formant data distributions for each phone are used to obtain the parameters of the sensory models, that is, the probability distributions over acoustic input for each phone category (term $P(I_j^t \mid [FeS_j^t = f])$), and more precisely, their parameters, i.e., the means and covariances of the Gaussian distributions for each phone in the lexicon (see Figure 3.2, black dots and black ellipses). In the case of the end-of-sequence marker #, it is arbitrarily mapped with formants normally distributed around the origin of the 2D formant space; such an arbitrary value is well outside of the meaningful formant descriptions for the vowels and the consonants of the lexicon, and thus silence is "easily recognized".

2.3 Phone duration and loudness profiles

In the current simulations, we consider that all phones have a constant duration of 50 ms, that is, 50 "time steps", except for initial vowels which have a duration of 100 ms (we keep the description of simulations in terms of time steps in the following, acknowledging that they would correspond to ms for application to real acoustic inputs). Nevertheless, syllables have variable duration since they have a variable number of phones. This number varies from 1 to 4: 1 for a non-terminal syllable made of a single vowel (e.g., the initial syllable in word *"apata"*), 2 if the vowel is followed by a final silence # (e.g., in the monosyllabic word *"a"*), 3 for a CV syllable with a phone for C and a phone for V connected by a transitional phone @, and 4 in a CV syllable that ends a word, because of the end-of-sequence phone. Accordingly, the duration of each word stimulus is displayed in column 4 in Table 3.1). For example, the word *"pata"* is composed of 50 ms of the phone *p*, followed by 50 ms of the transitional phone @, followed by 50 ms of the phone *a*, and so on, to end with 50 ms of the "end of word" marker #.

In addition to its description in terms of the temporal sequence of phones, each syllable is characterized by a loudness profile L which provides the input to the temporal control module for syllable segmentation. Loudness represents the auditory evaluation of acoustic intensity at a given time, resulting from the sensory processing of the acoustic signal envelope. This can be seen as capturing the variations of energy of the acoustic signal. In our simulations, loudness values are normalized between 0 and 1. Positive values of the local derivative of loudness are used to define onset events.

The loudness profiles used in this study are simplified, and serve to illustrate and capture the fact that there are syllabic energy fluctuations in real speech with, generally, rapid increase at syllable onsets and gradual decrease towards syllable offsets (in-between, almost anything can happen). In Figure 3.3, we display examples of loudness profiles we use in the simulations, respectively for the mono-syllabic word "*a*", composed of one vowel (Figure 3.3, left), for the bi-syllabic word "*pata*", composed of 2 CV syllables (Figure 3.3, middle) and for the tri-syllabic word "*apata*" composed of 3 syllables (one V and 2 CV syllables, Figure 3.3, right).

2.4 Paradigms for test conditions

We explored various test conditions for the model, in order to assess and illustrate the interaction between the bottom-up onset detection and the top-down onset prediction mechanisms, with the stimuli configured as presented above.

First, we consider a "nominal condition", in which the stimulus presents no difficulty, that is to say, the stimuli loudness profiles are "smooth and regular",

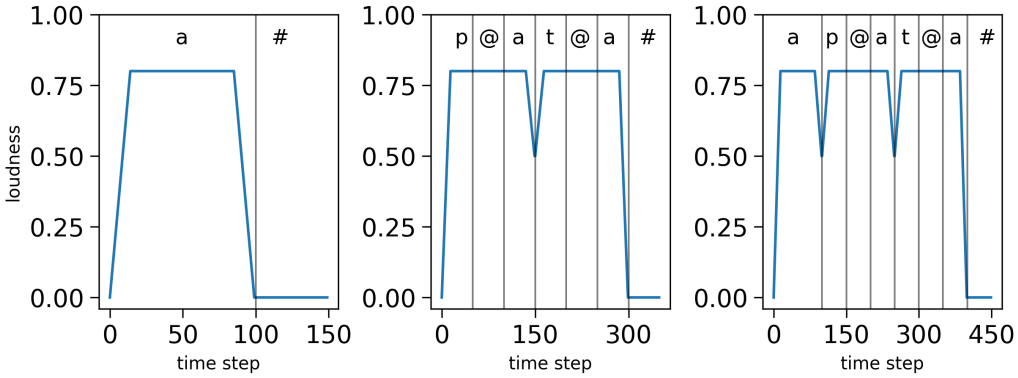

Figure 3.3: Examples of loudness variations for three input sequences: the word "*a*" (left), the word "*pata*" (middle) and the word "*apata*" (right). Simulated time is on the x-axis, and normalized loudness (arbitrary units) is on the y-axis. The vertical bars and top annotations refer to the associated phonetic content of the stimulus.

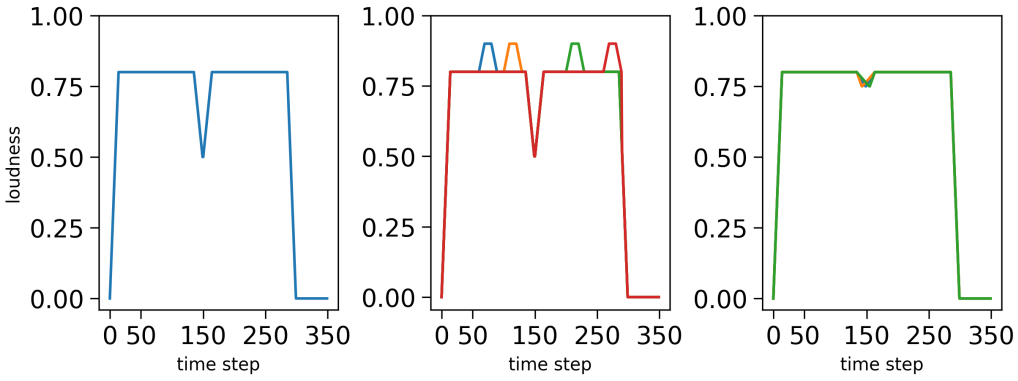

Figure 3.4: Loudness profiles for the bi-syllabic word "*pata*" used in the three simulation conditions: the nominal condition (left), the "noisy-event" condition (middle) and the "hypo-articulation-event" condition (right). Simulated time is on the x-axis, and normalized loudness (arbitrary units) is on the y-axis. Degraded conditions (middle and right) respectively correspond to the first noise level with one spurious event in a random position (with different positions corresponding to different colors) in the "noisy-event" condition, and to a dip height at 0.75 with random dip shapes (with different shapes corresponding to different colors) in the "hypo-articulation-event" condition.

such as shown in Figure 3.3.

Second, to assess the model in more difficult situations, we define degraded versions of the loudness profiles, in two possible ways. In the first case, we add noise events to the loudness profile, randomly positioned in portions where loudness is sustained in the nominal case: this may lead to the detection of spurious loudness events by sensory processing ("noisy-event condition"). In the second case, we decrease the depth of the loudness dip found at syllable boundaries and randomly modify the shape of the loudness dip: this may lead

sensory processing to miss syllabic onsets from the loudness signal ("hypo-articulation-event condition"). The three conditions, and the corresponding loudness profiles employed, are illustrated Figure 3.4 for the bi-syllabic word "*pata*".

2.5 Simulation configuration

We performed a set of simulations to evaluate the performance of the COSMO-Onset model. To do so, we simulated word recognition by the different model variants, for different words, and for the various test conditions; for the test conditions that simulate degradation of the stimulus, we applied different severity levels of the degradation. We now detail each of these components of our simulation set.

To recall, there are three considered variants of the model, in which syllable onset events are either assessed from bottom-up sensory information only (the "BU-only" model, in the following), or with top-down onset prediction combined with the *AND* operator (*AND* model), or, finally, with top-down onset prediction combined with the *OR* operator (*OR* model).

2.5.1 Degraded stimuli simulations

In degraded simulations, the stimuli we used for the experiment are all non-monosyllabic words from the lexicon (21 different words out of the 28 in the lexicon, see Table 3.1). Monosyllabic words were not used as stimuli since they would only contain a single onset event, at the initial iteration; nevertheless, they are part of the lexicon and are evaluated as possible candidates by the model during word recognition. Each of these words is presented once to the three variant models in nominal test conditions (i.e., with nominal loudness profiles).

In the "noisy-event" test condition, we considered 5 possible severity levels, by varying the number of noisy events applied to the loudness profile, from 0 (identical to the nominal case) to 4. Each noisy event lasts 10 % of the duration of the word, and its position is randomly drawn in the loudness profile of the word, ensuring that, when there are several noisy events, they do not overlap (see examples of severity level 1 on Figure 3.4, middle plot).

In the "hypo-articulation" test condition, we considered 5 possible severity levels, by varying the depth of the loudness dip between syllables. In this dip, loudness decreases to a varying minimal value, from 0.6 (identical to the nominal case) to 0.8 (in which case the loudness dip between syllables is entirely removed since loudness is at 0.8 inside syllables). The 5 possible values, therefore, are 0.6, 0.65, 0.7, 0.75, and 0.8. To introduce some variability, we randomly draw the precise time iteration, during the loudness dip, at which the minimum value is

attained (except, of course, for perturbation level 0.8, since the dip is removed altogether. See examples of a dip modified to be at value 0.75 on Figure 3.4, right plot)).

Note that, while severity level 0 of the "noisy-event" test condition perfectly corresponds to the nominal case (and the simulations are thus not repeated), this is not the case for severity level 0 of the "hypo-articulation" test condition, since the time instant of the loudness minimal value is varied, which may affect onset detection. Whenever perturbations would be randomly generated, we performed 10 independent simulations for that condition. Overall, we, therefore, performed 21*3*(1 + 4*10 + 4*10 + 1) = 5,166 word recognition simulations: 21 word stimuli, 3 model variants, 1 for the nominal condition, 4*10 for the noisy-event condition, 4*10+1 for the hypo-articulation condition.

2.5.2 Temporal misalignment simulation

We then assessed the robustness of the "BU-only" 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 the syllabic decoders. In other words, the "BU-only" model would compute onset detection in a normal fashion (term $P(OBU^t \mid \Delta L^t)$), but its output would be temporally delayed before being used in the computation of the inference for onset detection (term $P(OC^t = True)$).

We 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.

2.6 Performance measures

In order to evaluate the performance of the model variants during the simulations of word recognition, we use two performance measures: the unit identity performance metric (correct word recognition probability per se) and the boundary performance metric (correct onset detection per se). Both are already explained in Section 5.1 of Chapter 1.

However, the margins used in the boundary performance metric remain to be defined: we consider an event to be correctly predicted if the model generated an onset event internally in a 30-iteration wide time-window around the onset position in the stimulus (15 iterations before, 15 iterations after).

3 Results

We now report simulation results, to assess the performance of the three model variants in the three experimental conditions: the "nominal" condition, the "noisy-event" condition, and the "hypo-articulation-event" condition. First, we detail an illustrative example, allowing us to investigate the mathematical behavior of the model. This illustrative example is based on the input word *"pata"* in the nominal condition. Second, for each degradation condition, we first show the model behavior for the same input word *"pata"*, to illustrate mechanisms, before proceeding to the systematic evaluation of performance over the whole simulation set. And third, we show the results in the temporal misalignment condition for all the words of the lexicon.

3.1 Illustrative example in nominal condition

Figure 3.5 shows the simulation of the full model with the *AND* fusion in the nominal condition, for the example stimulus word "*pata*". It shows probability distributions computed by the model. Figure 3.5 (left) shows the different onset probability values in the temporal control module, and their evolution over time: the top-down onset prediction, the bottom-up onset detection composed of the refractory period and sensory event detection, and finally, the combined result with the *AND* fusion model. Figure 3.5 (right) shows probability distributions in the decoding module, with probability distributions over words (which provide the final output of the model), over syllables, and over phones.

Bottom-up onset detection shows that the model, based on sensory processing of the loudness envelope alone, would detect 2 events (Figure 3.5, left, bottom green curve), respectively around iterations 0, and 150. These, indeed, correspond to an increase in the loudness profile for the stimulus word "*pata*" (see Figure 3.4, left). Since these are outside the refractory period (dashed orange curve, left column, middle row of Figure 3.4), these two onset events are maintained and "output" by the bottom-up branch of the temporal control module.

Top-down lexical knowledge would predict 3 onset events (Figure 3.5, left, top red curve), respectively around iterations 0, 150, and 300. The first two match with bottom-up onset detection, and, since we illustrate here the *AND* fusion model, they are maintained in the output of the temporal control module (orange curve in the middle of Figure 3.5). The third onset predicted by top-down knowledge is due to the fact that the presented stimulus, the word "*pata*" is a prefix of other words in the lexicon (tri-syllabic words "*patata*" and "*patati*"). At this stage of word recognition, these three words are equally probable (see Figure 3.5, top right plot), so a third onset is likely. In this example, it is not confirmed by the bottom-up sensory event detection, and the *AND* fusion model

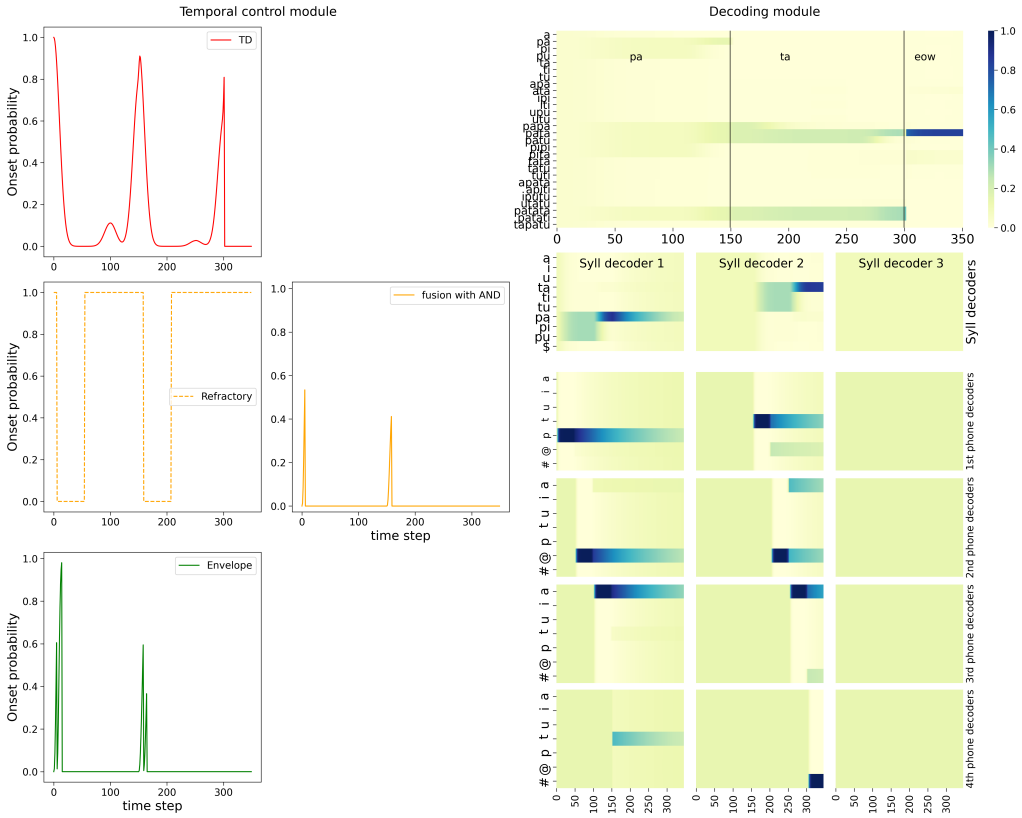

Figure 3.5: Example of simulation of the full model with the *AND* fusion in the nominal condition, on input word *"pata"*. Plots are organized to roughly map with corresponding positions in the model schema, see Figure 3.1. (Left, "Temporal control module" panel) Plot of the onset detection probabilities computed in the model (y-axis) as a function of simulated time (x-axis). Left column: In red, top-down onset prediction, in dashed orange, probability of an onset being outside a refractory period, in green, onset probability based on sensory processing of stimulus loudness. Middle plot: in orange, the onset probability of the *AND* fusion model (Right, "Decoding module" panel) Top plot shows the probability (color-coded, see the color bar on the right) over words (y-axis) as a function of time (x-axis). The vertical black bars and annotations at the top of the plot recall the stimulus structure; in this example, the stimulus is the word *"pata"*, with the acoustic signal of the first syllable during the first 150 iterations, the one of the second syllable during the next 150 iterations, followed by silence. Second row: plots of the probabilities (color coded) over syllables (y-axis), as a function of time (x-axis) during the activation of the corresponding syllable decoder. Bottom four rows: plots of the probabilities (color coded) over phones (y-axis), as a function of time (x-axis). Plots for phone decoders are sorted vertically, with the first phone above and the fourth at the bottom.

filters it out.

At each detected onset, the model activates a new syllable decoder, so we observe that 2 syllabic decoders are involved in the model (Figure 3.5, bottom right portion). In each syllable decoder, the probability distributions over syllables

evolve as acoustic input is processed, and the probability value of the correct syllable, that is, the one in the input, converges towards high values. We thus observe that each syllable decoder recognizes the appropriate syllable, which are /pa/ for the first syllable, and /ta/ for the second one. In the first syllable decoder, we observe a perfect competition for the first 100 time steps between all syllables beginning with phone *p*, and this competition gets disambiguated when phone *a* is processed. The second syllable decoder is activated around iteration 150 (it is a uniform probability distribution before activation), and shows a similar dynamic: first, competition between all syllables starting with phone *t*, then recognition of the correct syllable /ta/. The third syllable decoder is never activated, and thus remains uniform during the whole simulation.

Within every syllable decoder, phonetic decoders get activated sequentially (Figure 3.5, bottom 12 plots of right portion). We observe a behavior similar to the syllable decoders, except at a smaller timescale. Phone decoders stay uniform until their activation (this is especially visible for the phone decoders of the third syllable, which are never activated), then they decode the input, yielding, in this simulation, correct phone recognition, and after another onset is detected and predicted, the probability distributions gradually decay (this is especially visible for the phone decoders in the first syllable).

The probability distributions over syllables are then used, in the rest of the probabilistic computations in the model, to infer the probability distribution over words (Figure 3.5, top of the decoding module panel). Since syllable parsing was successful, so that syllable decoding was, too, then word recognition proceeds as expected, to recognize the word according to its syllables. Indeed, we observe that, at time step 150, that is, after decoding the first syllable /pa/, all words of the lexicon that start with /pa/ are equally probable. At time step 300, the lexical competition continues, and three words remain equally probable: the correct word "*pata*", and two competitors, the words "*patata*" and "*patati*", which embed the word "*pata*". This issue has been discussed in the literature (M. H. Davis et al., 1998; M. H. Davis et al., 2002); in the current illustrative simulations we do not address this general question, as it is naturally solved since we only consider isolated words: after a few iterations in which the acoustic input represents silence, the recognized word is the correct one, the word "*pata*".

We, therefore, observe correct onset detection (thus correct syllable parsing), but also correct phone, syllable, and word recognition by the full model with *AND* fusion. Simulating the model in either the "BU-Only" or the *OR* fusion variant, in the nominal condition, also provides correct answers and thus, good performance (simulations not shown here, see below for model performance evaluation), with the exception of the activation of a third syllabic decoder, when the top-down model relies on the *OR* model because the word "*pata*" is a prefix of other words

in the lexicon (this is not shown here but can be observed in the final simulation, in the "hypo-articulation-event" condition: see Figure 3.8).

3.2 Noisy-event condition

A first challenge for the listener is when the acoustic signal is perturbed, because for instance of external noisy conditions. In that case, the speech envelope can be degraded, introducing extraneous fluctuations of loudness and leading to detecting spurious events in the sensory processing of loudness. In other words, such spurious onsets would be detected by the bottom-up onset mechanism. Therefore, in this second simulation, we expect the "BU-only" model to result in erroneous syllable parsing, leading to incorrect syllable and word recognition. On the other hand, the complete model would rely on top-down lexical predictions of onset to "filter out" the unexpectedly detected onset (with the *AND* operator), leading to correct parsing and recognition.

Figure 3.6 shows the simulation of the "BU-only" variant of the model and of the full model (with the *AND* fusion model), on input word *"pata"*, with a degraded loudness profile that includes 2 spurious noise-events. The simulation we selected here for illustration adds these events at iterations 60 and 200 (see Figure 3.4, middle). We observe that with the bottom-up onset mechanism alone (Figure 3.6, top row), the bottom-up onset mechanism "fires" 4 events, corresponding to the 4 energy rises in the loudness profiles: near the start, then at iterations 60, 150, and 210. These are outside the refractory period of 50 ms, which would have otherwise filtered out these spurious onset events. Therefore, the bottom-up portion of the model detects 4 onset events. It leads to premature onset detection, which has a number of deleterious effects. First, it prematurely "closes" the first syllabic decoder, which was only fed with phone *p*, so that it is unable to correctly identify the first syllable in the input. Instead, the first syllabic decoder remains in an unresolved state of competition between all syllables that start with consonant phone *p*. Second, it prematurely opens the second syllabic decoder, which interprets the *a* vowel phone in the input as the syllable /a/, even though it is not legal in our lexicon in non-initial positions. This does not help resolve competition at the word level. Third, it correctly detects the real onset at time step 150 and opens a third syllable decoder supposed to decode the second syllable starting with phone *t*. But this is misaligned with the structure of the word "*pata*" which is bi-syllabic. Finally, the third decoder is prematurely closed, by the detection of the spurious onset event, near iteration 210. Overall, from one spurious event to another, the error in syllable parsing persists during decoding, and the bottom-up only variant is unable to correctly recognize the input word.

Compare with the simulation of the full model, with the *AND* fusion model,

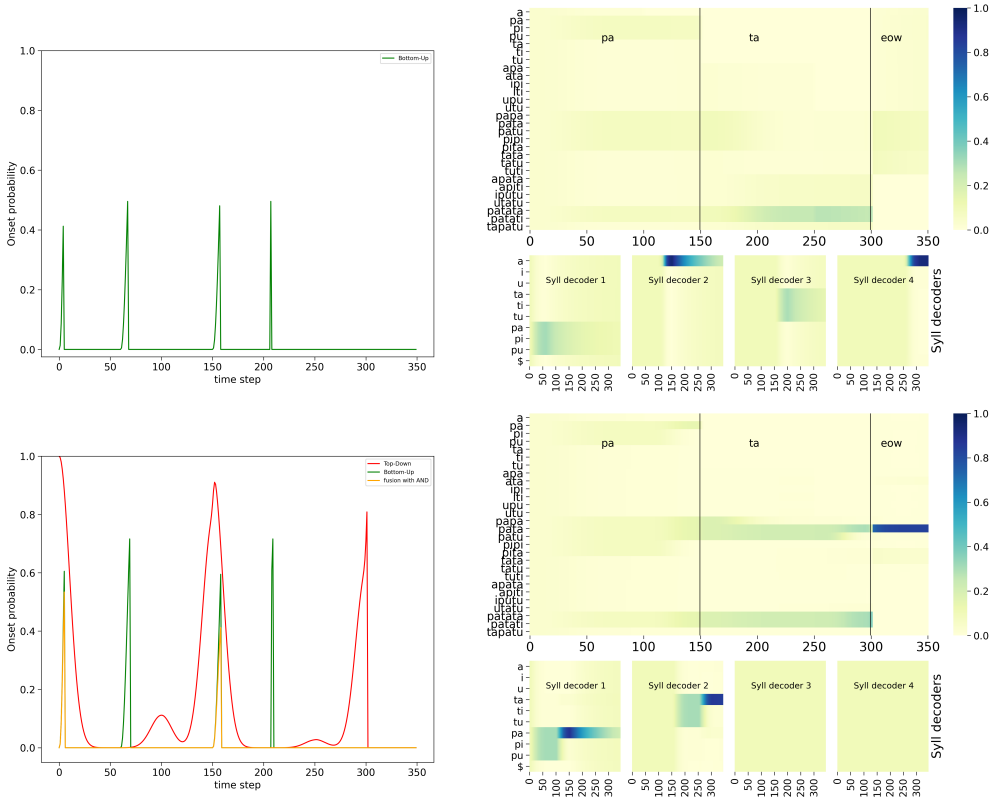

Figure 3.6: Example of simulation of the "BU-only" variant (top row) and the full model, with the *AND* fusion model (bottom row), in the noisy-event condition, on input word *"pata"*. Left column: plots of onset detection probabilities; Right column: plots of word probabilities in the word decoder (top row) and syllable probabilities of the syllable decoders (bottom row). Graphical content is presented in the same manner as in Figure 3.5 (except that onset probabilities are superposed in a single plot and the phone decoders are not shown).

on the same stimulus (Figure 3.6, bottom row). We observe that, while the bottom-up onset detection mechanism would lead to propose an onset near time step 60, the top-down temporal prediction model does not confirm this proposal. Therefore, the *AND* fusion model results in filtering out this event. This also happens with the other spurious event near time step 210. Therefore, with the *AND* fusion model, only the two "real" onsets are detected, that is to say, the ones at the start of each syllable. As a consequence, the behavior of the *AND* fusion model in the "noisy-event" condition is quite the same as in the nominal condition, with correct syllabic parsing, phone recognition, syllable recognition, and word recognition.

Figure 3.7 shows performance measures for the three variant models in the "noisy-event" condition, across all simulations. We first observe that both performance measures are highly correlated, suggesting that correct event detection relates to correct word recognition. Second, when there is no perturbation

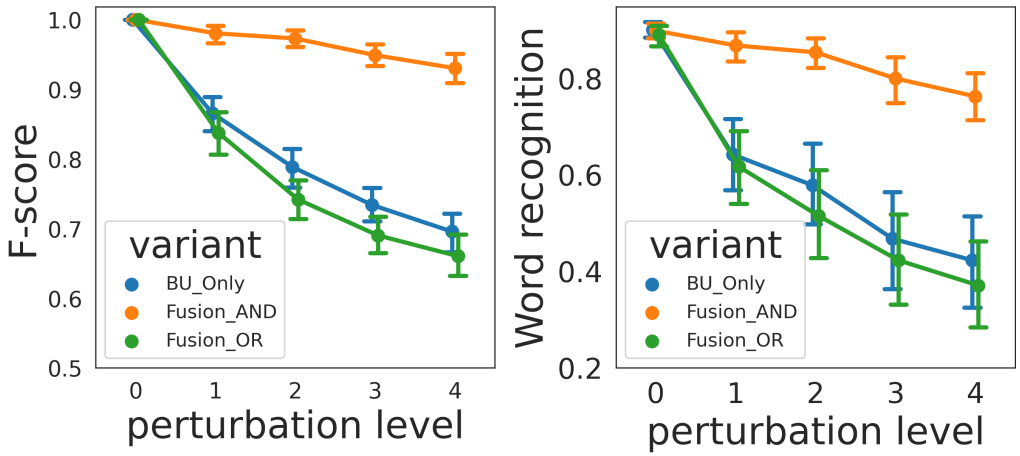

Figure 3.7: Performance of the three variant models in the "noisy-event" condition. Left: F-score (y-axis) as a function of the severity of degradation (x-axis). Right: word recognition probability (y-axis) as a function of the severity of degradation (x-axis). Every data point is averaged, over 21 words, and, where applicable, over 10 independent simulations with different randomly drawn perturbations.

(perturbation level 0), all variant models have the same performance, which is expected since top-down event prediction is redundant in this case with events that can be detected from the input signal. Third and finally, we also observe that the higher the severity level of degradation, the more performance decreases. Indeed, as degradation increases, the chance of having noise perturbations outside refractory periods increases, thus leading to more chances for spurious onset events. However, we observe that the model with the *AND* fusion is the most robust, as its performance decreases less with perturbation.

3.3 Hypo-articulation-event condition

In the second challenge we consider, degradation of the loudness profile leads to "removing out" onset events, for instance with an external perturbation masking a dip in acoustic energy at the syllabic boundary, or with this dip being much smaller, maybe because of hypo-articulation, or an error in speech planning, or excessive speed in speech articulation leading to speech slurring, etc. In that condition, we expect the "BU-only" variant of the model to miss onset events, leading to incorrect syllabic parsing, thus incorrect recognition. On the other hand, the complete model, with the *OR* operator, would use the lexically predicted onsets to insert them where the sensory onsets were missed, leading to correct parsing and recognition.

Figure 3.8 shows the simulation of the "BU-only" variant of the model and of the full model (with the *OR* fusion model), on input word *"pata"*, with the degraded loudness profile that decreases the dip depth in acoustic loudness at the syllabic boundary (see Figure 3.4, right). We observe that the "BU-only"

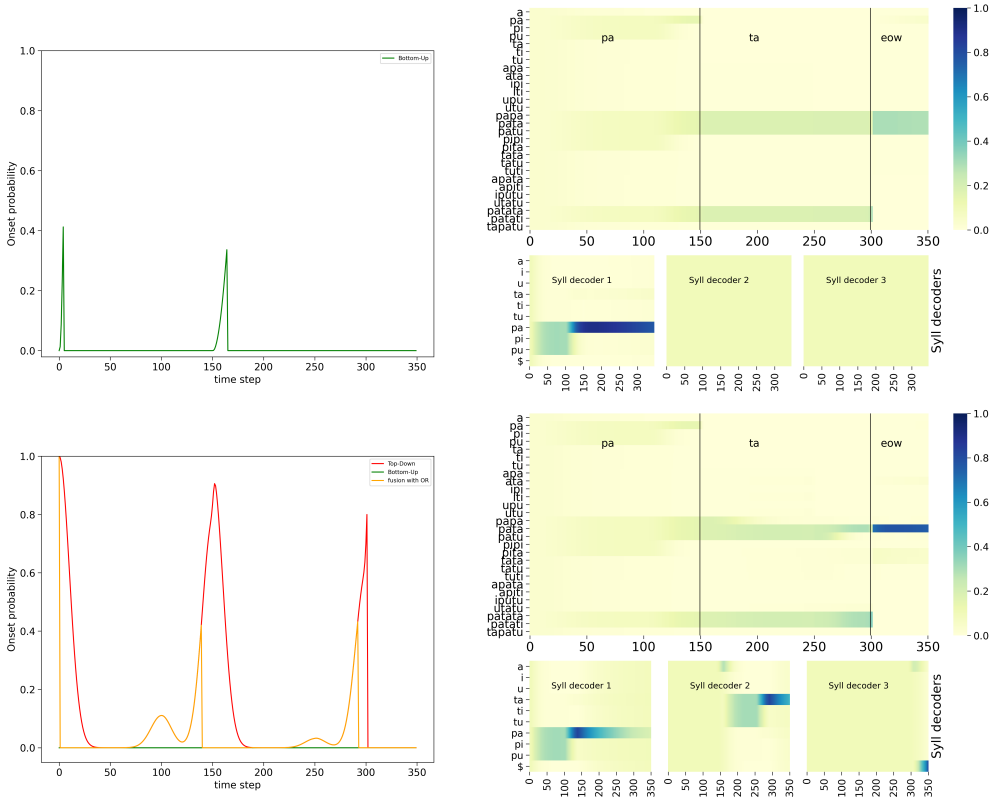

Figure 3.8: Example of simulation of the "BU-only" variant (top row) and the full model, with the *OR* fusion model (bottom row), in the hypo-articulation-event condition, on input word *"pata"*. Graphical content is presented in the same manner as in Figure 3.6.

variant of the model does not ascribe a probability for the second onset prediction that is high enough (the probability is lower than the decision threshold at 0.4), and therefore it misses the second onset (near time step 150) so that the first syllabic decoder stays activated for too long. Although it correctly recognizes the initial /pa/ syllable, it never activates the second syllable decoder. This leads to unresolved competition at the word level between all the bi-syllabic words starting with syllable /pa/, and, ultimately, incorrect word recognition.

In contrast, with the *OR* fusion model (Figure 3.8, bottom row), the top-down onset prediction allows recovering the missed onset event at time step 150, which helps to avoid the problem of faulty syllabic parsing and misalignment of the second syllabic decoder with the stimulus. In this condition, the full model with the *OR* fusion model leads to correct syllabic parsing, phone recognition, syllable recognition, and word recognition. Notice, however, that simulations here are not exactly the same as those of the model in the nominal condition, with two differences that merit attention.

Indeed, we first observe that the syllabic decoders are a few iterations ahead

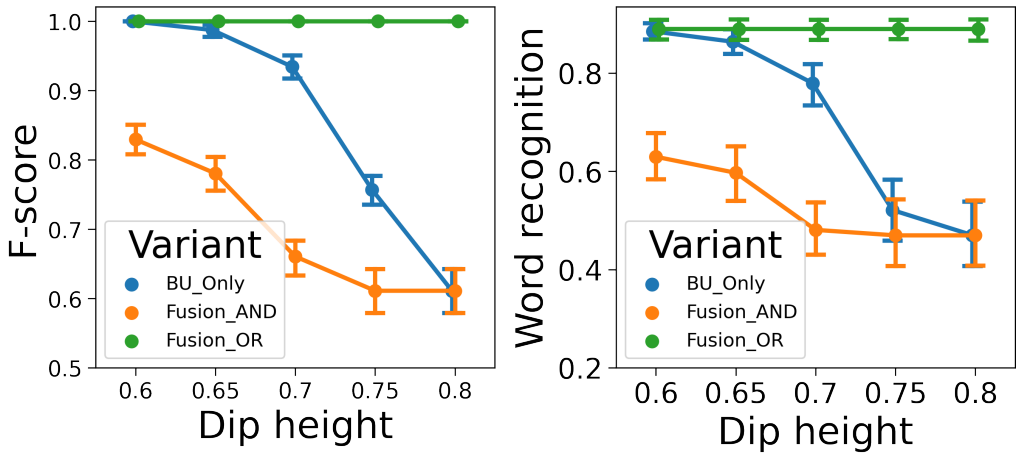

Figure 3.9: Performance of the three variant models in the "hypo-articulation-event" condition. Left: F-score (y-axis) as a function of the severity of degradation (x-axis). Right: word recognition probability (y-axis) as a function of the severity of degradation (x-axis). The severity of degradation is measured with the minimal value attained in the loudness dip at syllable boundaries ("dip height" in plot labels; 0.6 corresponds to a well-marked dip, 0.8 removes the loudness dip altogether). Every data point is averaged over 21 words, and over 10 independent simulations with different randomly drawn perturbations.

of the stimulus: for instance, whereas the syllabic boundary between the first and second syllables is exactly at time step 150 in the stimulus, the first onset event resulting from the *OR* fusion model is around time step 140. This leads the second syllabic decoder to process, for a few iterations, the end of the first phone *a* in stimulus *"pata"*. This slight temporal misalignment is due to the value we set for the onset decision criterion, at 0.4. Such a value is reached early of the "bump" in onset probability provided by the lexical model, which correctly peaks at time step 150 (Figure 3.8, bottom left plot, compare the lexical prediction (red curve) and output of the *OR* fusion model (orange curve)).

The second notable behavior in this simulation is the activation of a third syllabic decoder. Indeed, the lexical onset prediction model is aware of words in the lexicon which embed "*pata*". Therefore, up to time step 300, there is an unresolved competition, at the word recognition level, between the embedding words containing "*pata*" and "*pata*" itself. A third syllable could then, from the lexical prediction, be expected, so that an onset event is lexically generated. This leads to activating a third syllable decoder, which mostly processes the "end of word" marker in the acoustic input (after the few iterations where it processes the end of the second *a*, because of the slight temporal misalignment discussed above). Observing a third syllable "composed of silence" is only consistent with the word *"pata"* in the lexicon, so that it is, ultimately, correctly recognized.

Figure 3.9 shows performance measures for the three variant models in the

"hypo-articulation-event" condition, across all simulations. First, we observe, here again, that both performance measures correlate. Second, we observe that, contrary to simulations in the "noisy-event" condition, all three model variants do not have the same performance for the less degraded condition. Indeed, in our simulation, we randomly select the time iteration at which the minimum value is reached; this changes the geometry of the dip in loudness, so that, even though it has nominal depth (when the dip height is 0.6), it can misalign onset detection and prediction, which negatively affects the *AND* fusion model. Third and finally, we also observe that performance decreases as degradation increases, for the "BU-only" and the *AND* fusion models. The performance of the *OR* fusion model, on the contrary, does not decrease as perturbation increases, indicating the robustness of the *OR* fusion model in the "hypo-articulation-event" condition.

3.4 Temporal misalignment condition

Figure 3.10 shows the experimental results in which the manually-inserted delay is varied systematically. We observe an overall inverted-U shaped plot (top 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 also analyzed results independently for monosyllabic, bisyllabic, and trisyllabic words (bottom plots). We observe that monosyllabic words are overall better recognized, and performance is more robust (longer plateau for varying delays); this, of course, is due to the fact that monosyllabic word recognition is only dependent on a single syllabic onset detection. However, result patterns for bisyllabic and trisyllabic words are very similar to the global results.

Overall, these experimental results suggest that, when the "BU-only" model processes the simplified stimuli that we have defined, 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.

4 Discussion

This chapter enabled us to present the first implementation of the COSMO-Onset model where the **X** and **Y** portions of the decoding module and the temporal control module, respectively, are designed to process simplified speech stimuli. Of course, the simulations we proposed in this chapter are mostly illustrative and certainly preliminary. They should be extended, in various directions that we discuss now, and that will be partly explored in the next chapters.

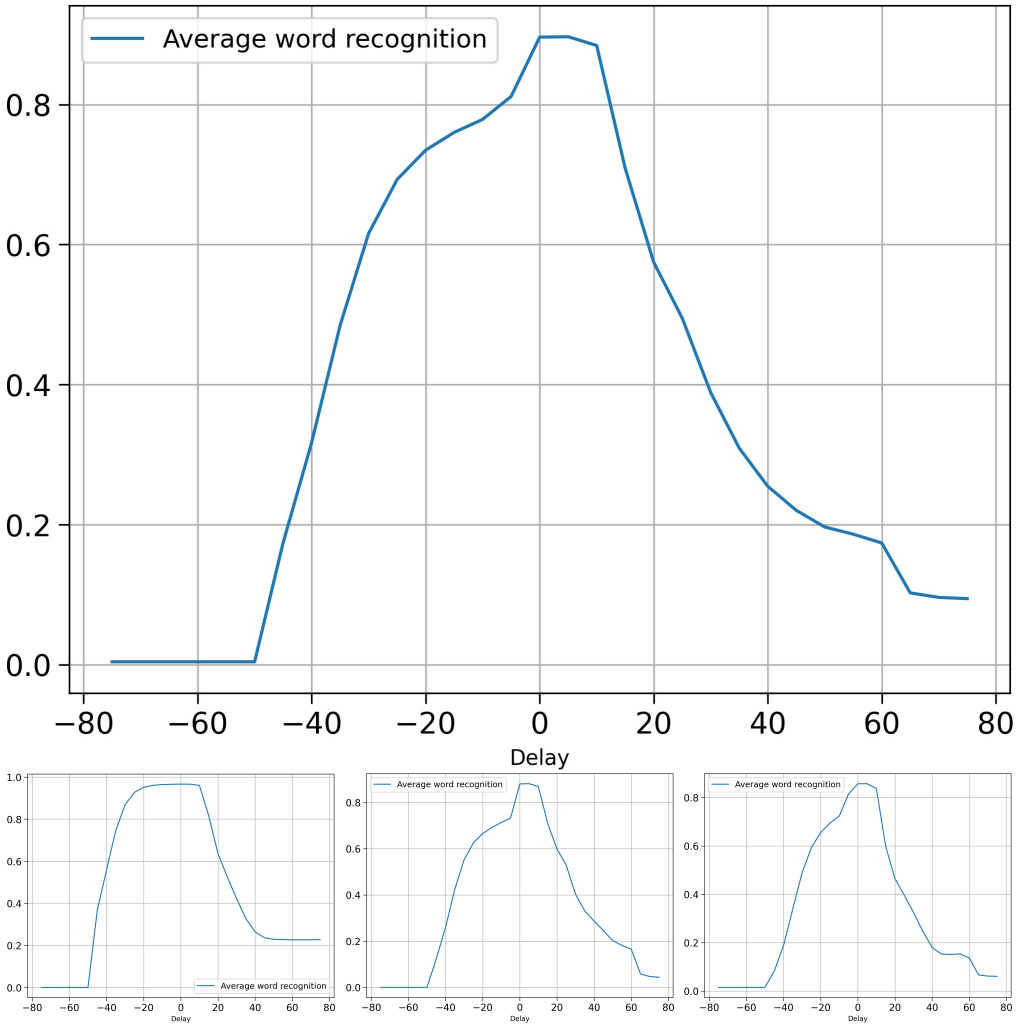

Figure 3.10: Temporal misalignment experiment. Top plot: average probability for the correct word (y-axis) in the "BU-only" model variant simulating word recognition, overall words of the lexicon, as a function of a manually-imposed delay between onset detection and their use for opening and closing the syllabic decoders (x-axis, in iterations). Bottom plots: the same as above, but for monosyllabic (left), bi-syllabic (middle), and tri-syllabic (right) words only.

Efficiently combining bottom-up and top-down information for syllabic parsing

The simulations presented in Section 3 suggest a rather clear overall picture. Firstly, in the "hypo-articulation simulation set", generating missing events in the bottom-up branch, the *OR* model behaves efficiently and outperforms the "BU-Only" model in terms of both detection accuracy and recognition score. Secondly, in the "noisy-simulation set", generating spurious events, the *AND* model discards most of these spurious events and outperforms the "BU-Only" model once again in terms of both detection accuracy and recognition score.

Notice that in both cases, the bottom-up branch performs actually better than the non-adapted fusion model. Indeed, the *AND* model degrades event detection when it is already difficult in the hypo-articulation case, probably because of a slight asynchrony between the bottom-up and the top-down branches; and the *OR* model slightly increases the number of inaccurate or spurious events detected in the noisy case, probably because the top-down information enhances spurious envelope modulations.

Globally, this raises the question of selecting the right model for the right stimulus condition. This falls into the general question of model selection and averaging, for which literature is abundant (e.g., Burnham & Anderson, 2004; Wasserman, 2000). This would suggest various ways of analyzing the probabilistic content of each of the three models "BU-Only", *AND*, and *OR* and selecting or averaging their output accordingly. Importantly, the rationale of the two sets of simulations suggests that some exogenous contextual criterion could be used for model selection. Thus, if the system is able to extract some evaluation of the level of noise or the quality of articulation during a short period of time, this information could be used as a proxy to select the *AND* or the *OR* fusion model accordingly or even to combine them. The same kind of endogenous or exogenous information could also be used as a prior or weight in the Bayesian fusion process involved in both the *AND* and the *OR* model. For example, instantaneous estimates of the noise level could act as a weighing factor in the *AND* Bayesian fusion process, increasing/decreasing the respective roles of the bottom-up and top-down branches accordingly. The Bayesian framework that we have adopted all along this work in the development of the COSMO-Onset model is obviously adapted to study and explore all the corresponding questions about model selection and fusion.

Finally, if there indeed exist two different fusion modes, namely an *AND* and an *OR* behavior, this raises some interesting questions for cognitive neuroscience, asking whether specific neural markers could be associated with a shift from one mode to the other. Indeed, it has been proposed, for instance by Giraud and Poeppel (2012) or Arnal and Giraud (2012), that there could exist specific frequency channels respectively associated with bottom-up (theta channel) and top-down (beta channel) messages. The shift from the *AND* to the *OR* behavior, possibly associated with noisy conditions vs. hypo-articulation, would result in different coordination in time between theta and beta bursts, that could be explored with adequate neurophysiological paradigms.

From "toy" stimuli to realistic speech processing

First of all, it is important to acknowledge that the current material used as input to the model is far from real speech. To be able to finely monitor the model

output at this preliminary stage, we designed toy stimuli.

This is first the case for the synthetic loudness curves used to simulate the speech envelope and the two kinds of adverse perturbations applied to these curves, together with the simplified loudness processing in the bottom-up branch performing a "simplistic" bottom-up onset detection with straightforward envelope analysis. A further step in the development of the presented implementation of COSMO-Onset in this chapter, will be to consider a more realistic neuro-computational model able to track the signal envelope and adapt in an online manner to variations in instantaneous syllabic frequency, as in oscillatory models such as the ones developed by Hovsepyan et al. (2020), Hyafil, Fontolan, et al. (2015) or Räsänen et al. (2018) (see also the neuro-physiological refinements recently introduced by Pittman-Polletta et al. (2021)). Importantly, these various existing models should help provide COSMO-Onset with a possible neurophysiological implementation of the temporal processing component of the algorithmic structure presented on Figure 3.1 in the **Y** box, which would make the relationships between the present simulations and real neurophysiological data more straightforward. As already mentioned in the literature review chapter (Chapter 1), the oscillatory model developed by Räsänen et al. (2018) seems the simplest and the most concerned only with segmentation processes. Thus it is a good candidate to replace the bottom-up onset detection mechanism implemented in the **Y** box in this chapter. This will be discussed in Chapter 4.

Secondly, the spectral description of the acoustic stimulus was limited to the first two formants. The first layer in the COSMO-Onset implementation presented in this chapter (Figure 3.1), that is the Phone Sensory Layer, currently takes for granted the feature extraction from the speech input by directly implementing phone recognition from the first two formants, while realistic spectral analysis of speech utterances would rather exploit a bank of auditory filters (e.g., gammatones (Hohmann, 2002; Patterson et al., 1992) or Mel-cepstrum analysis (Rabiner, 1989). The latter spectral analysis has received large interest in the Automatic Speech Recognition (ASR) domain and has shown to be efficient in encoding the essential features of the speech spectral content. Thus, it is a good candidate for that component of COSMO-Onset. We can therefore replace all the current mechanisms implemented in the **X** box with a more sophisticated model to directly perform syllable recognition from the speech input, of course after extracting the features such as the Mel Frequency Cepstrum Coefficients (MFCC). This will be the main focus of Chapter 5.

From there on, the ability of COSMO-Onset to deal with specific experimental conditions displaying the role of top-down processes in syllabic parsing and onset detection can be tested with more realistic experimental settings. For this aim, we target the experimental data by Aubanel and Schwartz (2020) showing that

natural speech is more intelligible in noise than speech rendered isochronous, while isochrony also plays a role in helping intelligibility, but to a lesser extent. Naturalness and isochrony play here complementary roles which could fit quite well with the existence of a bottom-up onset detection branch exploiting isochrony, and a top-down prediction branch exploiting naturalness. Capitalizing on a second implementation of COSMO-Onset equipped with more realistic architectures for the temporal control module (Chapter 4) and the decoding module (Chapter 5), we will attempt to partly replicate these data in Chapter 6.

Chapter 4

A study of the oscillation-based syllabic segmentation model by Räsänen et al. (2018)

Note

This chapter is adapted from a published conference paper (Nabé, Diard, et al., 2022).

In Chapter 2, we presented the COSMO-Onset model with its conceptual architecture (Figure 2.1). To recall, there are two components that we introduced in a generic form in the conceptual model, the first concerned with syllable decoding (noted **X**, the light blue box of Figure 2.1), and the second with bottom-up onset detection (noted **Y**, the dark blue box of Figure 2.1). We then developed an implementation of this conceptual model suited for simplified stimuli in Chapter 3 (Figure 3.1), in which the **X** box was replaced by phone processing layers that inform syllable decoding, whereas the **Y** box was based on the processing of the loudness of the speech signal looking for rapid increases.

In this chapter, we start moving towards a version of COSMO-Onset able to process natural speech input. Here, we focus on the **Y** box, replacing the simplistic bottom-up onset detection mechanism implemented in the illustrative version of COSMO-Onset with a more realistic onset detection process inspired by neural oscillations. As discussed in the literature review, among the various oscillatory models, the model developed by Räsänen et al. (2018) stands out as the simplest, operating on simple mechanisms of envelope detection and linear second-order oscillators. In contrast to other discussed models, this model is solely concerned with acoustic envelope processing in order to detect syllabic onsets. In the remainder of this chapter, we refer to this model as the ***RDF***

model after its authors [1].

The present contribution has four major objectives. Firstly, we apply the ***RDF*** model on a French corpus, the *Fharvard* corpus developed by Aubanel et al. (2020), that we have already presented in Section 6 of Chapter 1. This will enable widening its evaluation set for syllabic onset detection. Secondly, we will extend the ***RDF*** model to the detection of P-centers on the same French corpus, to assess its performance relative to the nature of syllabic events. Thirdly and crucially in the present context, since our key experimental paradigm is concerned with the role of isochrony in the bottom-up processing of natural speech (as presented in Section 6 of Chapter 1), we assess whether isochrony plays a role in the efficiency of the ***RDF*** model. Finally, and still, in the context of our target experimental paradigm, we will evaluate the model's robustness to noise.

1 Simulation Material

1.1 Corpus

The present simulations exploit the acoustic *Fharvard* corpus from Aubanel et al. (2020), introduced in Chapter 1. From the 700 sentences of the *Fharvard* corpus, in this study, we only used a subset of the overall dataset that was fully annotated by the authors at various levels, namely at the word, syllable, P-center, and phoneme levels. This subset amounts to 177 sentences composed of multi-syllabic words with a total of 646 distinct syllables. Figure 4.1 shows the distribution of syllable duration in the corpus.

An important characteristic of the sentences in this corpus concerns their relation with isochrony. To evaluate the role of isochrony in event detection, we characterize each sentence by its distortion to isochrony as defined in Chapter 1, Section 6. It is computed for the time series of syllabic events t (which can either be syllabic onsets or P-centers) by the distortion to transform the natural time series into a target time series t' made of the same number of isochronous events, thanks to the formula:

$$\delta = \sqrt{\frac{\sum_{i=1}^{N} (\log \tau_i)^2 d_i}{\sum_{i=1}^{N} d_i}} \ ,$$

with d_i and d'_i respectively the duration between successive events in the reference and target times series ($d_i = t_{i+1} - t_i$; $d'_i = t'_{i+1} - t'_i$), and τ_i the time-scale factor between the reference and target time series: $\tau_i = d'_i/d_i$.

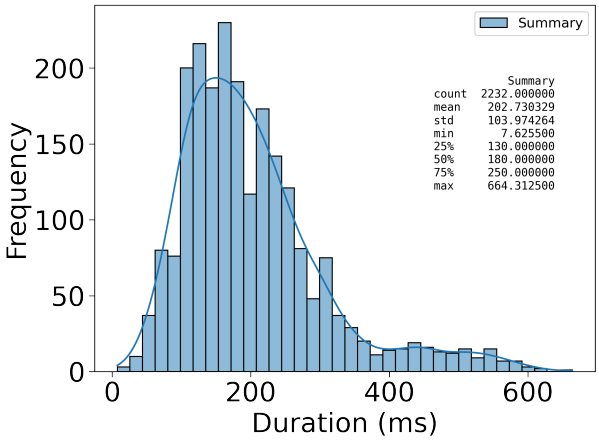

Figure 4.1: Histogram of syllable duration in the *Fharvard* corpus (Aubanel et al., 2020) (mean $\sim$ 203 ms, median $\sim$ 180 ms, mode $\sim$ 154 ms).

1.2 The *RDF* model

To recall, Figure 4.2 shows the different processing stages sequentially performed by the ***RDF*** model. The original acoustic waveform is first processed by a Gammatone filter-bank resulting in outputs $e_c(t)$ (each channel with its separate color), which are then transformed by a linear oscillator into oscillator amplitudes $x_c(t)$, that are finally combined to obtain the sonority envelope $S(t)$, as the model output. This output can then be used to drive the search for local extrema.

Initially, the authors used the ***RDF*** model to detect only the "valleys" (troughs) in the sonority envelope, which are considered the syllable boundaries. They correspond to the red bars of the last bottom plot on Figure 4.2.

We then extended this initial version to detect P-centers. Let us recall that P-centers correspond to the "psychological moment of occurrence" of syllables (Morton et al., 1976; Strauß & Schwartz, 2017). Even if this definition is consensual, a precise acoustical landmark that would correspond to P-centers is still lacking. However, they are classically determined by determining peaks of energy increase of the speech envelope, or in our case, the sonority envelope (Marcus, 1981; Patel et al., 1999). Typically, this is easily found by analyzing the first-order derivative of the sonority envelope. Therefore, we extended the ***RDF*** model to detect peaks in the first-order derivative of the sonority envelope, which is the oscillatory model output.

To implement the extension, we used the Python version of the ***RDF*** model, which we adapted accordingly in the last step. Instead of the sonority envelope being returned as the output of the model, we take its first derivative to search for local maxima. In the following, we will interpret these local maxima as events

[1]There exists an implementation of the ***RDF*** model mainly in *Matlab* that can be found in this GitHub repository. Also, a Python implementation can be found here.

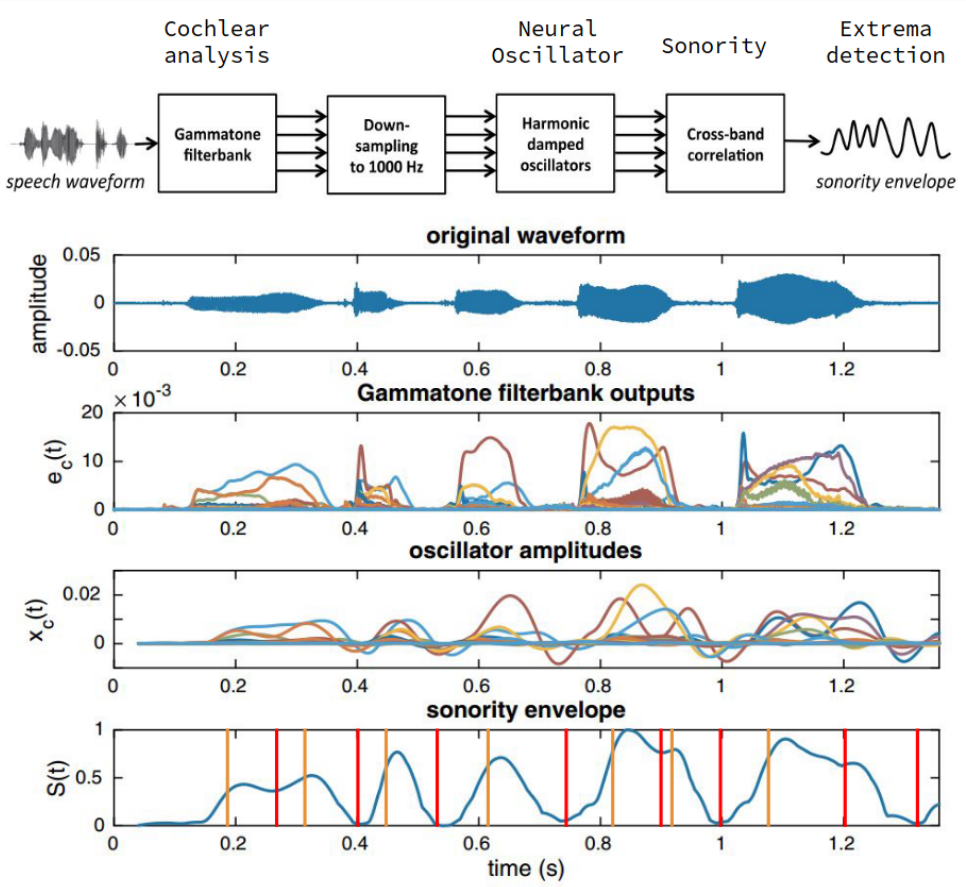

Figure 4.2: An example of the processing stages for an utterance "Can you get the circle?" by the ***RDF*** model. From top to bottom: 1) the different processing stages in the ***RDF*** model transforming the speech input into the sonority envelope (model output), 2) the waveform of the input sentence, 3) the different Gammatone filter-bank outputs applied to the input signal, 4) the amplitudes of oscillator resonance applied to each Gammatone filter-bank, and 5) the sonority envelope obtained by a linear combination of the oscillator amplitudes annotated with the temporal events: the red bars correspond to syllable boundaries, and the orange bars correspond to P-centers.

corresponding to the detection of syllable P-centers. They correspond to the orange bars of the last bottom plot on Figure 4.2.

For our performance measure, we use the boundary detection metric with a margin of 50 ms. This is identical to the measure used in the initial study of the ***RDF*** model (Räsänen et al., 2018), which will allow comparison with our results. Finally, to evaluate the model's robustness to noise, we added white Gaussian noise to the initial speech data, with varying signal-to-noise (SNR) ratio from -30 dB (very noisy) to 30 dB (almost noise-free) by steps of 10 dB (totaling 7 SNR values).

1.3 Parameter calibration

The ***RDF*** model has four free parameters, that require calibration to ensure optimal use of the oscillator algorithm. To obtain optimal values for each of these parameters, we performed a search on a predefined grid of values as in the original paper by Räsänen et al. (2018). To perform calibration, we optimized performance on a training dataset with 100 audio files within the 177 available ones, while all experimental results provided below were obtained from the remaining 77 sentences in the test set. We now recall the model parameters and define our 4-dimensional calibration grid.

The first parameter is the central frequency f_0, that is, the resonant frequency of the oscillator, which varies in the theta frequency band, and usually depends on the speaker and the speaking conditions. We considered 7 calibration values, from 5 to 8 Hz with a .5 Hz step.

The second parameter is the quality factor Q, a function of the central frequency and the bandwidth of the oscillator $Q = f_0/\Delta f$. It measures the damping rate of the oscillator. A notable value is $Q = .5$, for which the oscillator is critically damped so that it would follow the envelope of the signal as closely as possible. For larger values of Q ($Q > .5$), corresponding to an under-damped oscillator situation, the oscillator resonates more around its central frequency, with a slower decay of its amplitude, even if it is no longer excited by a real signal. For smaller values of Q ($Q < .5$), corresponding to an over-damped oscillator situation, the oscillator performs more temporal smoothing, with less dependence on its central frequency. We considered, for calibration, an empirically defined set of 21 possible values for parameter Q: .15, .25, .5, .75, from 1 to 1.9 with a .1 step, and from 2 to 5 with a .5 step.

The third parameter is the minimum detection threshold *thr*, that is, the minimal difference between a local extremum and neighbor extrema enabling us to consider the local extremum as meaningful. We considered 3 possible threshold values: .01, .025, and .5.

The fourth parameter is a fixed delay *del*, to shift all detected events, so as to mitigate artifacts introduced by signal processing techniques, in particular, delays due to smoothing, filtering and windowing operations. For syllabic boundary detection, we considered 15 possible values (0 to 70 ms with a step of 5 ms); for P-center detection, we considered 7 possible values (0 to 30 ms, step of 5 ms).

Table 4.1: Parameter values resulting from calibration on the training set, and resulting F-scores on the test set.

	P-centers	Syllable boundaries
f_0 (Hz)	6.5	7
Q	1.4	1.9
thr	0.025	0.01
del (ms)	0	55
F-score	.89	.75

2 Simulation Results

2.1 Performance on syllabic event detection in French

Table 4.1 provides the optimal values, resulting from calibration, for model parameters, for both P-center and syllable boundary detection tasks. To recall, calibration was performed on the training set. On Figure 4.3, we observe that the best performing parameters, both for P-center and syllable boundary detection, correspond to Q factor values of an under-damped oscillator. It also displays a varying distribution of performance values ranging from almost 0.7 to 0.89 for P-center detection, and from almost 0.35 to 0.75 for syllable boundary detection. We also observe that the optimal f_0 value at 7 Hz for syllable boundary detection is higher than the inverse value of the mean syllable duration (mean syllable duration is 203 ms, the inverse is 4.9 Hz). This was also the case for all simulations in the original study (Räsänen et al., 2018). However, the optimal f_0 value seems to be close to the inverse of the mode or of the median of the asymmetric distribution of syllable duration (see Figure 4.1), which suggests that such statistics could better describe the overall speech rate in the corpus, with respect to the ***RDF*** oscillatory model.

Table 4.1 also reports detection performance on the test set for both tasks, using the optimal parameter values. Performance for syllable boundary detection is measured by an overall F-score of .75, which is comparable to previous experimental results in Finnish, Estonian, and English (Räsänen et al., 2018). In contrast, performance is quite higher for P-center detection, with an overall F-score of .89.

2.2 Role of isochrony in event detection

2.2.1 Relation between isochrony in the distribution of syllabic boundaries and P-centers

Figure 4.4 shows the correlation between lack of isochrony for syllabic boundaries and P-centers, for all sentences in the experimental corpus. We observe that there is no significant correlation: sentences with low distortion to synchrony

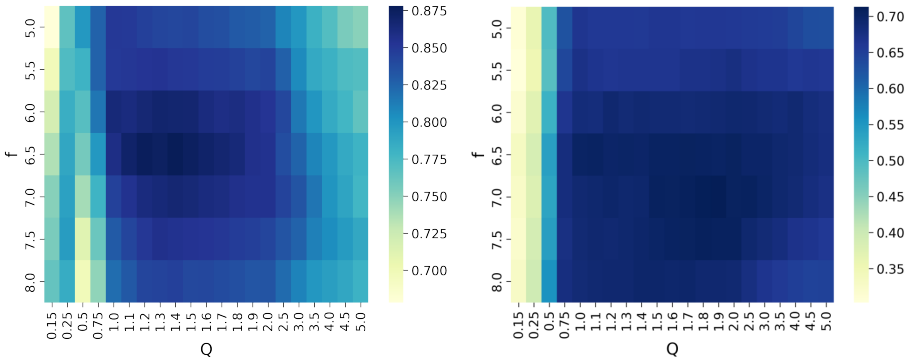

Figure 4.3: Model performance (color value) as a function of central frequency f_0 (on the y-axis) and Q factor (on the x-axis). On both the left plot (P-center detection) and the right plot (syllable boundary detection), the darker the color, the higher the model performance.

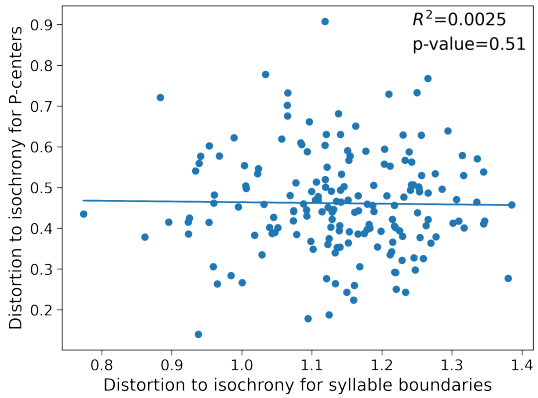

Figure 4.4: Correlation between distortion to isochrony values δ computed with respect to syllabic boundaries (x-axis) and P-centers (y-axis), for the 177 sentences of the experimental corpus. Linear regression (solid line) and corresponding squared correlation coefficient R^2 are indicated in the plot.

in syllabic boundaries may have large distortion for P-centers, and vice-versa (Pearson correlation coefficient $R = 0.05$, p-value $p = 0.51$). In the following, we use only distortion to synchrony computed over the distribution of P-centers, in line with the experimental study by Aubanel and Schwartz (2020).

2.2.2 Relation between distortion to P-center isochrony and event detection

Figure 4.5 shows the variations of event detection performance as a function of distortion to P-center isochrony, for P-center detection (left) and syllable boundary detection (right). We observe that for both P-center and syllable boundary detection, there is a statistically significant negative correlation between

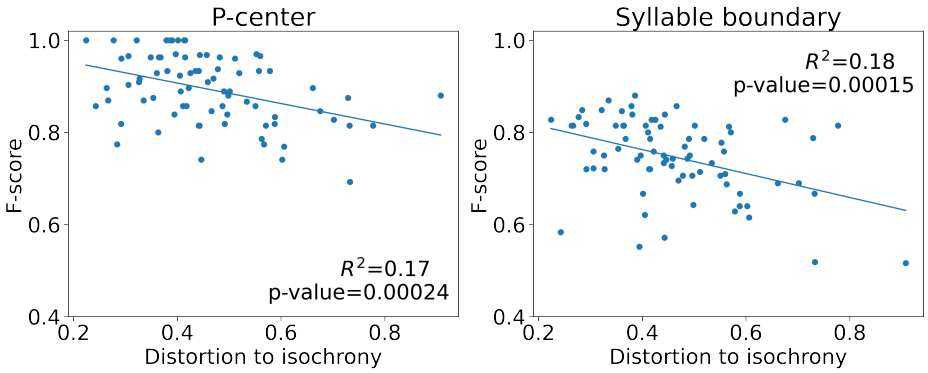

Figure 4.5: Event detection performance (F-scores, y-axis) against P-center temporal distortion to isochrony (δ, x-axis), for P-center detection (left) and syllable boundary detection (right). Linear regressions (solid lines) and corresponding squared correlation coefficients R^2 are indicated in the plots.

model performance and temporal distortion. In other words, model performance is higher, and events are better identified, when temporal distortion is small, that is to say, for natural sentences which happen to be more isochronous.

2.2.3 Role of the resonance factor in event detection

Figure 4.6 shows event detection performance as a function of the Q factor when all other model parameters are fixed, for P-centers (left) or syllable boundaries (right). Strikingly, the best performance is obtained for resonant systems with Q values much larger than the so-called critical damping value $Q = .5$ which corresponds to a system that essentially tracks the acoustic envelope with no additional resonance process. While the optimal value for the Q factor is similar for P-centers and syllable boundaries in the $1.2 - 1.5$ range, the adequate range is rather restricted for P-centers, with quasi-optimal values between 1.1 and 1.8 and then a rapid decrease for too resonant systems; in contrast, a large range of Q values above 0.75 are adequate for syllable boundary detection, although detection performance is lower overall.

2.3 Event detection in noise

Figure 4.7 shows how model performance varies as a function of the signal-to-noise ratio (SNR). The ***RDF*** model appears to be rather robust to noise, with its performance almost unchanged up to a rather large level of noise (SNR at 0 dB), with performance sharply decreasing for lower values of SNR.

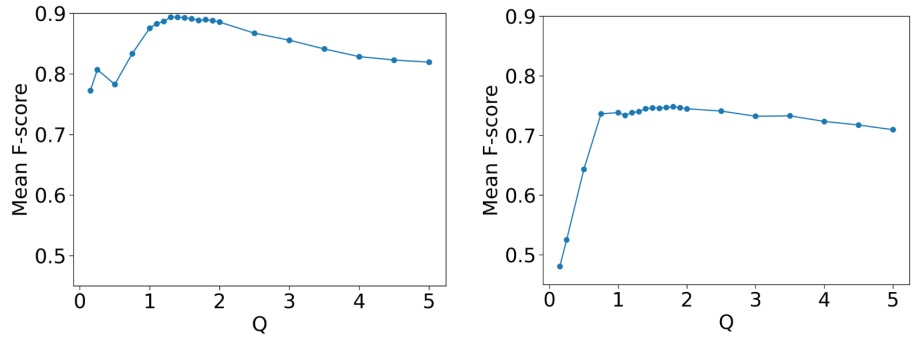

Figure 4.6: Event detection performance (mean F-scores, y-axis) against the Q parameter value (x-axis), for P-center detection (left) and syllable boundary detection (right).

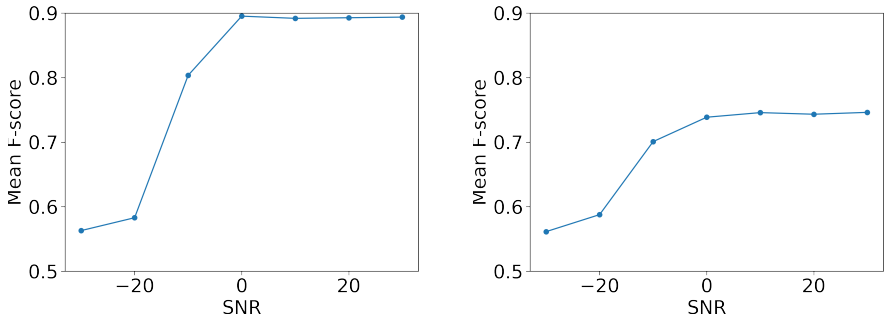

Figure 4.7: Model performance (mean F-score, y-axis) with respect to the noise level (SNR, x-axis) for P-center detection (left) and syllable boundary detection (right).

3 Discussion

In this chapter, we have evaluated the ***RDF*** oscillatory model of event detection (Räsänen et al., 2018) on a French corpus, and shown that it performs as well as previous evaluations on other languages.

Importantly, the results point to the role of resonance mechanisms in this process. Indeed, it appears that (1) the system performs better for resonant than for non-resonant characteristics of the proposed algorithm (see Figure 4.6) and (2) acoustic speech signals with higher inter-P-center isochrony lead to better event detection (see Figure 4.5). Furthermore, the detection process based on a resonant response to envelope modulations appears more efficient to detect P-centers than syllabic onsets (see Table 4.1). This is likely because P-centers are more robust events within the speech envelope dynamics. It could lead to proposing segmentation algorithms involving P-center detection as a complement signal to syllable boundary detection: although P-centers are not systematically related to syllable onsets (see Figure 4.4), P-center detection is a likely signal

that a syllable boundary preceded, and was possibly missed.

The event detection system of the ***RDF*** model appears rather robust in acoustic noise. Still, our study, in line with results from previous experiments, suggests that performance is far from perfect (with 25 % missed events for syllabic onsets and 11 % for P-centers) without noise, and rapidly degraded for noise at SNR values under 0 dB. This suggests a potential role for top-down processes, exploiting statistics of sentence rhythms in relation to lexical, syntactic, and prosodic knowledge. In their study on comprehension of speech in noise, Aubanel and Schwartz (2020) showed that, while natural isochrony improved comprehension, anisochronous speech re-timed to become more isochronous is actually less well perceived, which points to the role of top-down predictive processes in speech segmentation. This is the core of the COSMO-Onset model we have previously discussed in the last chapter (Nabé et al., 2021) to model how bottom-up and top-down information could be combined for speech syllabic segmentation. The present study provides an important baseline: the ***RDF*** model is a purely bottom-up, signal-driven event detection model. Our objective remains to study how complementing it with top-down knowledge could improve the overall syllabic event detection performance.

Chapter 5

A syllable recognition model using Random Forests

In the previous chapter, we started developing an implementation of COSMO-Onset capable of dealing with real speech input by studying an oscillatory model of speech segmentation for implementing the bottom-up onset detection process (the dark blue box **Y**) presented in Chapter 2. The other component that needs to be adapted is the **X** component associated with syllable decoding (the light blue box).

In this chapter, we focus on this **X** box, replacing the syllable decoding mechanism dealing with simplistic speech features implemented in the illustrative version of COSMO-Onset with a more elaborated algorithm operating on more realistic speech features. For this, we use a machine learning algorithm called "Random Forest" (RF) for the syllable decoding part, operating on classical speech features that are Mel Frequency Cepstral Coefficients (MFCC).

In the following, we present the RF algorithm and justify its use for our decoding problem. Then, we carefully present all the material required to implement a decoding module based on RF. Finally, we present the model performance evaluation results.

1 A Machine Learning algorithm: Random Forest

In order to build our syllable recognition model, and since the target data set is the one from Aubanel and Schwartz (2020), we need a machine learning model (Bishop & Nasrabadi, 2006; Mitchell & Mitchell, 1997) able to achieve good categorization performance on small data sets. Therefore this discards deep learning methods from our consideration (Bengio, 1993; Goodfellow et al., 2016; LeCun et al., 2015) since their performance largely relies on the amount of available data for the learning process (Halevy et al., 2009; Sun et al., 2017).

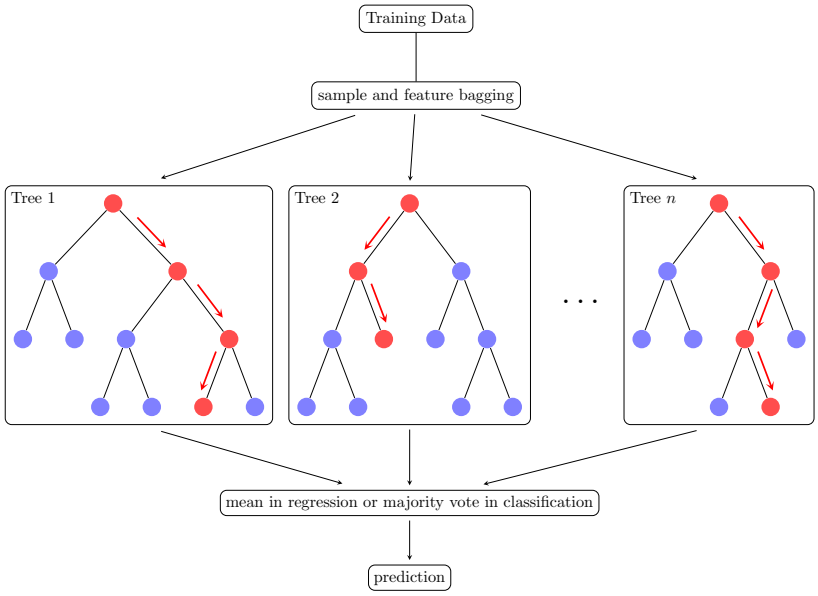

Figure 5.1: Illustration of the random forest algorithm. A random forest with n trees, where each tree has a different number of nodes (blue and red circles, with the red circles being the path undertaken in each tree).

There are various machine learning algorithms, including parametric models (which assume a particular data distribution) and non-parametric models (which do not assume any specific data distribution and are thus usually more versatile). Another distinction separates supervised and unsupervised learning algorithms, depending on whether their training data is labeled or not, respectively.

The random forest algorithms (Amit & Geman, 1997; Biau & Scornet, 2016; Breiman, 2001) are part of the supervised non-parametric models and belong to a more general algorithmic class, called "ensemble methods" (Dietterich, 2000a; Sagi & Rokach, 2018; Zhou, 2012). These are models which build on other models. In a nutshell, instead of using only a single model to train on data and make predictions, a set of different models are constructed and aggregated to provide a final prediction. Typically, for a classification task, a set of classifiers are learned and a voting system is used for the final prediction, while for a regression task, a set of regressors would be computed, and the final prediction would be the average of their outputs (illustrated on Figure 5.1) [1]. This class of models is supposedly more robust than single classifiers and regressors and often leads to better prediction results (Huang et al., 2009).

Ensemble methods differ in their way of aggregating single predictors (also called weak learners). In a non-exhaustive way, we can mention "Bayesian averaging" (Domingos, 2000; H.-C. Kim & Ghahramani, 2012), "Bagging" (Breiman, 1996), and "Boosting" (Drucker et al., 1992; Schapire, 1999). In the "Bayesian

[1] Source: Code to generate the random forest illustration.

averaging" method, the probabilistic predictions of various models, weighted by their respective posterior probabilities, are linearly combined. The "Bagging" method (contraction of *Bootstrap* and *Aggregating*) is a general aggregation method that creates subsamples (bootstrap samples) from the original data set, builds a predictor from each sample, and makes decisions by averaging all individual predictors. The "Boosting" method differs from the "Bagging" method by the way it trains weak learners. In the latter, weak learners are trained independently from each other and in parallel, whereas in the former, weak learners are learned sequentially and adaptively to improve the overall model predictions. It is still unclear whether a particular version would be preferred over others, as various factors affect the performance of these methods. For example in situations with classification noise, the bagging method has been shown to be better than boosting (Dietterich, 2000b).

The random forest algorithm, as initially conceived by Breiman (2001), uses the bagging learning method by randomly splitting the original data set based on some criteria (e.g., number of features for each weak learner, number of samples for each weak learner, etc) and by training weak learners independently. Furthermore, as the name suggests, the weak learners of the random forest algorithm are decision trees (Breiman et al., 2017; Kotsiantis, 2013; Rokach & Maimon, 2005).

Decision trees logically combine a series of elementary tests in a sequential manner. In the case of numeric features, each test compares a numeric feature against a certain threshold. In contrast, in the case of nominal and/or categorical features, each test compares features against a set of possible values. Decision trees can be viewed as a sequence of conditional propositions. As such they are different from other models of machine learning considered black-box models since one can trace the model decision and interpret its output. This explainability property transfers from decision trees to random forests.

The random forest algorithm presents many advantages, from its robustness to over-fitting (thanks to the ensemble technique), to its versatility, to its high performance. Recently, a study by Grinsztajn et al. (2022) showed that decision tree-based models such as random forest outperform deep learning methods on tabular data (data sets with a fixed number of features). Prior to this study, (Fernández-Delgado et al., 2014) realized a large study where they evaluated 179 classifiers from 17 families of models on 121 data sets from the *UCI collection of datasets*, and concluded that the random forests algorithms were the best classifier among these families. Besides, the random forest algorithm has been used in ASR systems for various tasks. In language modeling, for instance, Xue and Zhao (2008) have used a random forest on top of an HMM model for phonetic acoustic modeling and showed that it achieved better performance on small data

sets. In audio-visual speech recognition, some studies have used combined visual and audio features to train random forest classifiers (Borde et al., 2020; Terissi et al., 2015), with Terissi et al. (2015) also showing it outperformed classical HMM models, especially in noisy conditions. Without an exhaustive list of the many applications of random forests, we also mention that they have been used in speaker recognition tasks (Nawas et al., 2021) or in isolated word recognition (Attar et al., 2010), where authors showed it outperformed classical HMM-based models of speech recognition.

In the rest of this chapter, we present the simulation material, how we developed our syllable recognition model based on random forests, and discuss performance results.

2 Simulation Materials

As in the previous chapter, the present simulations also exploit the acoustic *Fharvard* corpus from Aubanel et al. (2020), introduced in Section 6 of Chapter 1. Similarly, instead of using all 700 sentences of the corpus, we use the same subset of 177 sentences. To recall, Figure 4.1 shows the distributions of syllables duration. We also recall that this corpus of 177 sentences is thoroughly annotated by the authors at different linguistic levels, from the sentence to the word, syllable, and phoneme levels.

We used the material presented here to train and evaluate a random forest model of isolated syllable recognition. However, instead of aiming at recognizing the distinct syllable categories themselves, which is a difficult task to solve with our limited corpus, we develop a random forest model for recognizing syllabic structures. The reason for this choice is practical, as it allows defining broader categories, thus increasing the number of samples in each category. We now present the syllabic corpus itself.

2.1 Syllabic corpus

Figure 5.2 shows a block diagram of our procedure for syllabic corpus creation. The inputs are the 177 audio files (sentences) and their corresponding annotations, which include, among other annotations, the syllabic alignment annotations. Prior to any forward processing, we strip all audio files from silent segments at their beginnings and ends. Doing so assures that our database's audio content only contains speech signals. Then, we process both inputs simultaneously. First, for every audio file, we parse its syllabic annotation where syllable labels are located between two boundaries (the onset and offset of the syllable). The information is used to segment the audio file appropriately into all of its constituent syllables, except when it is the "silence" syllable, as it sometimes occurs within a sentence,

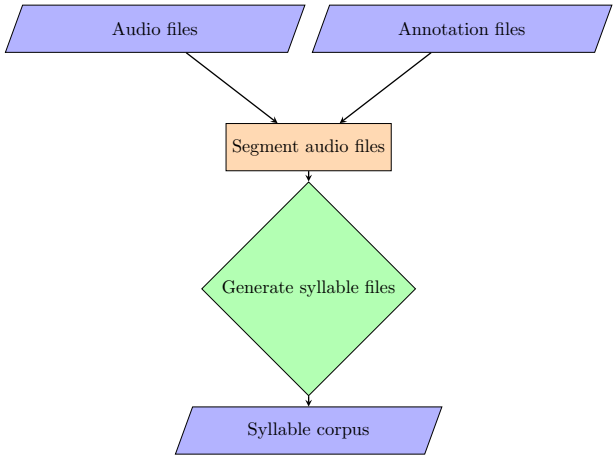

Figure 5.2: Block diagram of syllabic corpus creation. The first two top nodes in light blue represent the inputs. They are passed onto a processing stage (in light orange), which generates syllable files (the light green node). The last step corresponds to assembling all the files into the syllabic corpus (output).

and discard it instead. Next, we generate a syllable file for each audio segment, thus creating the complete syllable corpus.

With the syllable corpus generated in this manner, we obtain 2,232 syllables files (with an average of 12.6 syllables per sentence). For practical considerations (the exact reason will be discussed later), we filter out syllable files with a duration of less than 35 ms, removing 5 occurrences (one respectively for syllables **/œ̃/, /a/, /nə/, /o/, /y/**. This leaves us with 2,227 syllables files in the corpus, corresponding to different realizations of 645 unique syllable categories, with varying durations (see Figure 4.1).

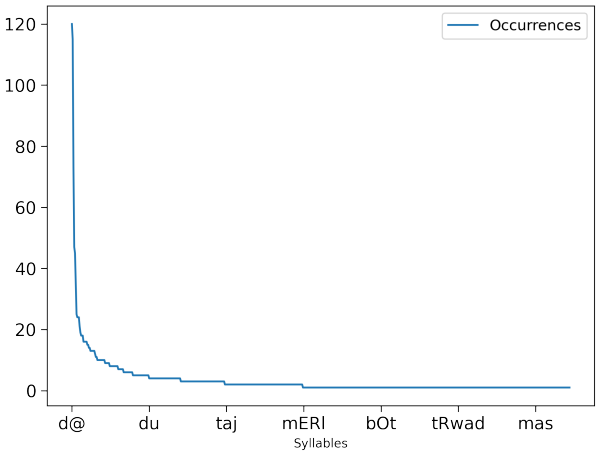

Figure 5.3: Histogram of syllable occurrence in the syllable corpus generated from the *Fharvard* corpus Aubanel et al., 2020 (mean $\sim$ 3.5 occurrences, median $\sim$ 1 occurrence). On the x-axis, is the syllable category, and on the y-axis, is the count of occurrences.

Table 5.1: Occurrence counts with respect to the 13 initial syllable types. Each syllable type (first column) is associated with the number of syllable categories composing it (second column), the total count of occurrences in the syllable corpus (third column), and an example of a syllable with the given structure (fourth column).

Types	# of syllable categories	# of occurrences	Example
V	10	108	/a/
VC	5	25	/il/
VCC	2	2	/ɛst/
CV	190	1372	/sa/
CVC	225	376	/syʁ/
CVCC	29	31	/vɛʁs/
CVCCC	2	2	/maʁbʁ/
CCV	103	206	/pli/
CCVC	61	78	/swaʁ/
CCVCC	4	4	/pʁɔpʁ/
CCCV	10	19	/tʁfii/
CCCVC	3	3	/spʁɛs/
C	1	1	/ʁ/
Total	645	2,227	

Figure 5.3 shows the syllable occurrence distribution for the 645 identified unique syllable categories. The top 25 % (the 75 % percentile) of the syllables have a minimal occurrence of 3. This also means that the vast majority of the syllables have at most only 3 occurrences. Furthermore, we can observe the effect of the outliers that drive the mean towards 3.5, although, in reality, many syllables are repeated only once in the whole corpus.

Thus, in the following, we consider the **syllable type** rather than the syllable category itself. The syllable type is of the form C^nVC^m, where C stands for *consonant* and V for *vowel*, n and m range from 0 to 3, and the superscript notation indicates repetition n or m times (as in a regular expression notation). Rather than building a model to recognize isolated syllable categories, we consider a model to recognize syllable types.

Table 5.1 contains global statistics on syllable distributions depending on their consonants and vowels composition. Initially, the database contains 13 distinct syllable types, which we further regroup into 6 composite categories. The first four syllable types we consider are the categories, CV, CVC, CCV, and CCVC, which are largely represented in our corpus. However, the remaining syllable type categories have fewer data samples, and we collapse categories further. First, we group syllable types V, VC, and VCC together into one syllable type that we hereby note as VC*. Second, all the remaining syllable types are grouped into a final syllable type that we note as "others". In the following, we use the term "syllable type" for these 6 general, composite syllable types.

Table 5.2: Occurrence counts with respect to the 6 composite syllable types. Each syllable type (left column) is associated with the number of syllable categories composing it (middle column) and the total count of occurrences in the syllable corpus (right column).

Types	# of syllable categories	# of occurrences
VC*	17	135
CV	190	1372
CVC	225	376
CCV	103	206
CCVC	61	78
others	49	60
Total	645	2,227

Table 5.2 shows the occurrence count of distinct syllable categories and data samples in the 6 composite syllable types. As expected, the syllables with the most complete structure (CV and CVC) account for most of our database. We also observe that the minimum number of samples is 60 ("others" syllable type), which is supposedly sufficient to train our model.

All in all, we have a syllable corpus of 2,227 data samples, distributed across 6 labels, corresponding to the 6 syllable types in which we chose to regroup the 645 syllable categories. [2]

2.2 Performance measures

For performance measures, since we are interested in the categorization model, we use the classical accuracy and recall measures on the model predictions. We also provide additional, combined performance measures, namely the precision and the F-score.

2.3 Building the Random Forest model

To provide an acoustic input to the model, we extract acoustic features and build a fixed-size vector that represents the audio file. Following Hovsepyan et al. (2020), we fix this size to $N = 8$ frames. As is done classically, we first extract the MFCC features for every frame of every audio file. This is done with the Librosa library (McFee et al., 2015), which is widespread in the speech and audio processing community. Classically, we extract 13 MFCC coefficients per 25 ms frame from 512-points FFT, with a number of samples between successive frames (*hop length*) fixed by the desired number of frames. For each audio input, the

[2] We remark that there is a single C syllable in the corpus. This is very likely an error in the annotation. It comes from the utterance "un sentier raide est pénible pour nos pauvres pieds" ("a steep path is hard on our poor feet" in English) which corresponds to example sentence number 526 in the *Fharvard* corpus. The error is located specifically in the annotation of the word "pauvres" which they split into the syllables /pov/ and /ʁ/.

hop length is then hence computed by:

$$hop\_length = \frac{Duration \times SR - n\_fft}{N - 1}$$

where $Duration$ is the audio signal duration, SR is the sampling rate equal to 22,050 kHz, N is the number of frames, and n_fft is the number of samples used to compute the Fast Fourier transforms.

This fixes the minimal duration of a speech signal segment to 35 ms to ensure the possibility to compute MFCC values without errors, for 8 different frames. This is the technical reason justifying that we do not consider segments shorter than 35 ms in our corpus.

Finally, each syllable audio file is represented by a vector of 104 features (8 frames of 13 MFCC coefficients), to which we add the duration of each audio file. Therefore, this results in a vector of 105 features provided as an input to the random forest model. The random forest algorithm is implemented using the Scikit-Learn library (Buitinck et al., 2013; Pedregosa et al., 2011) [3].

Parameter calibration of the random forest classifier was performed with a grid search over sets of values for each of the algorithm's parameters. To perform calibration, we optimized performance (model accuracy) on a training data set with 1,672 syllable audio files within the 2,227 files overall, while experimental results provided below were obtained from the remaining 555 audio files in the test data set. These data sets were obtained by applying a splitting method of 75 %-25 % of data samples in each syllable type, that is, for every syllable type, there are 75 % examples in the training data set, and the remaining in the test set.

The random forest model has several parameters that are considered key to its performance:

1. *n_estimators*: The number of decision trees in the classifier. Intuitively, the higher this number, the best it may overcome over-fitting. However, in practice, too many trees may lead to poor models. We explored values between 100 and 1000, with a step of 100. The best value we found during optimization was 700.

2. *max_features*: The maximum number of features for every decision tree. In our random forest, both data samples and the number of features are bootstrapped. Some decision trees may be trained on some features, and others on other features. We explored values between 5 and 10, with a step of 1. The best value was 9.

3. *min_samples_leaf*: The minimum number of samples required to be at a

[3] Link to scikit-learn random forest model.

Table 5.3: Performance scores of the random forest model on the test data set. Top rows: performance for each composite syllable type, sorted in decreasing F-score order. Bottom row: overall performance scores. Scores are given in columns, with recall, precision, and F-scores, respectively in the second, third, and fourth columns.

Performance (in %)	Recall	Precision	F-score
CV	93.6	82.1	87.5
VC*	84.8	82.35	83.6
CVC	64.9	64.9	64.9
CCV	37.25	63.3	46.9
CCVC	15.78	100	27.27
others	13.3	66.67	22.22
Overall	78.2	84.3	80.5

leaf node of a decision tree. We explored values between 4 and 10, with a step of 1. The best value was 7.

4. *min_samples_split*: The minimum number of samples required to split a node of a decision tree. We explored the same parameter range as the previous parameter, from 1 to 10, with a step of 1. The best value was 2.

5. *min_impurity_decrease*: A node will be divided if its division leads to a better information gain. This gain of information is characterized by the value of the *min_impurity_decrease* parameter. We explored the set of values $\{10^{-5}, 10^{-6}, 10^{-7}\}$. The best value was 10^{-7}.

6. *max_depth*: The maximum depth (level) of the decision trees. We explored values from 10 to 50, with a step of 10. The best parameter value obtained was 30.

7. *max_samples*: The fraction of the total number of samples in the training data to train each decision tree. We explored the values from 0.5 to 1 with a step of 0.1. The best parameter value obtained was 0.9.

3 Simulation Results

Table 5.3 shows the scores of the different performance measures of the random forest model on the 555 samples in the test data set. We observe that it achieves overall scores above 70 % for all performance measures. Particularly, it has higher precision than recall, resulting in a good F-score. Breaking down performance for each syllable type, we observe that the model is overall better, according to F-scores, for the CV, CVC, and VC* syllable types than for the remaining three syllable types. Finally, Table 5.4 provides a confusion matrix between the 6 composite syllable types, in which we verify that most mistakes can be

Table 5.4: Confusion matrix of the random forest model predictions for each syllable type on the test data set. Ground truth, the expected categories, are in rows, and columns are model predictions. For instance, out of the 343 test samples for the expected category CV, 321 samples were correctly recognized as CV, 5 samples were miscategorized as VC*, 9 as CVC and 8 as CCV.

	CV	VC*	CVC	CCV	CCVC	others
CV	321	5	9	8	0	0
VC*	5	28	0	0	0	0
CVC	29	1	61	3	0	0
CCV	30	0	1	19	0	1
CCVC	3	0	13	0	3	0
others	3	0	10	0	0	2

considered as "expected mistakes", that is to say, with confusions between close syllable types (for instance, mistaking CV for CVC or CCV, or mistaking CVC for CCVC).

4 Discussion

In this chapter, we have developed and studied a model of isolated syllable type recognition using a random forest method (Breiman, 2001). We evaluated the model performance on a subset of the French corpus *Fharvard* (Aubanel et al., 2020) used by Aubanel and Schwartz (2020). For a test set of 555 syllables, it provides reasonably good syllable-type-recognition scores by achieving more than 70 % on the classical performance measures for a categorization algorithm, that is to say, precision, recall, and F-score (see Table 5.3).

Globally, the resulting model provides a good balance of accuracy among the 6 composite syllable types (see Table 5.4), making errors that we can classify as reasonable since the confused syllable types are rather similar. Interestingly, we observe that there is a correlation between the syllable type complexity and the number of training samples. Logically, scores are higher when the number of training samples is larger (see Table 5.1). The poorest performance is reached for the "others" syllable type, the least well-represented type, and also arguably the more difficult one since it is a highly diverse "catch-all other syllable types" class. However, it is important to note that in any syllable type, performance, as measured by the F-score, is still better than random guessing, which would be, in our categorization task with 6 classes, at 16.67 %.

Since we have a somewhat limited corpus, with an unbalanced distribution of classes (the 6 syllable types have different training data set sizes), and since our goal is not to build a fully featured "state-of-the-art" speech recognition algorithm, we consider that the obtained random forest model we developed is satisfactory for our purposes. We, therefore, exploit this model in the remainder of our work,

as a syllable-type recognition mechanism to replace, in the COSMO-Onset model, the **X** box.

Chapter 6

COSMO-Onset: Adapting to real speech

Note

This chapter will provide the basis for a paper to be submitted to an international journal.

To quickly recall, in Chapter 2, we presented the COSMO-Onset model at the conceptual level, showcasing its principal components and leaving out two portions as unspecified (the **X** and **Y** portions), that we then defined later on, first in a version aiming at illustrating the model behavior (see Chapter 3) and second in a version aiming at dealing with real speech input. This was the focus of the last two chapters, with Chapter 4 describing how we defined the **Y** portion of the model with an adaptation of the ***RDF*** model, and Chapter 5 describing how we defined the **X** portion with a classification algorithm based on Random Forests.

In this chapter, we finalize the definition of the COSMO-Onset model. In the following, we present the different changes that need to be applied to go from the COSMO-Onset model presented in Chapter 2 to a version able to process real speech input. Then, we state our theoretical hypotheses and present the simulation material used to evaluate the model's behavior. Finally, we present results addressing our main theoretical hypotheses.

1 COSMO-Onset for real speech stimuli: putting it all together

Figure 6.1 shows the variant of COSMO-Onset for real speech. Even though this final version of COSMO-Onset is indeed able to process real speech stimuli from a

real corpus, we stay within the limitations already introduced in the two previous chapters. Indeed, since our focus is on evaluating whether our model is able to account for the experimental results provided by Aubanel and Schwartz (2020), all developments in the present chapter aim at treating the corpus of this experiment, which is limited in size. Therefore, as in the previous chapter, the present variant of COSMO-Onset is designed to perform only syllable type recognition. As a consequence, since it is impossible to directly relate syllable types with word identity, we consider a variant of the initial architecture of the model, without the word recognition level in the decoding module. The architecture of the temporal control module stays basically the same, even though, without the word recognition layer, we have to adapt the top-down temporal prediction accordingly.

1.1 Adapting the temporal control module

To recall, the temporal control module of the COSMO-Onset model is comprised of two interacting parts. Bottom-up onset detection relies on mechanisms operating only on the signal itself, namely the speech envelope, whereas top-down onset prediction is based on mechanisms involving "higher-level", linguistic knowledge, such as the constituent syllable duration of words.

1.1.1 Adapting the bottom-up onset detection

When presenting the COSMO-Onset model in Chapter 2, we designated by the **Y** box the bottom-up onset detection portion of the temporal control module. In this chapter, we replace it with the onset detection ***RDF*** model (Räsänen et al., 2018), that we introduced, extended, and studied in Chapter 4.

Even though they both operate on the speech envelope, the main differences with the bottom-up onset detection mechanism used in the first implementation of the COSMO-Onset model in Chapter 3 reside in the fact that the ***RDF*** model uses "real neural oscillatory" principles. This makes it a more realistic model capable of extracting useful information from speech dynamics, to detect syllable onset events with good performance.

In the first variant of the model, the bottom-up onset detection mechanism was based on using simplistic envelope differentiation in search of energy increases, characteristic of syllable onsets. However, with an oscillatory model such as the ***RDF*** model, this is done using more sophisticated signal processing mechanisms operating on the speech envelope that will, later on, drive event detection by either looking for troughs (syllable onsets) or "peaks" (syllable P-centers). Furthermore, in the first variant model, due to a lack of oscillatory mechanism, we used a complementary mechanism and defined a refractory period in order to avoid "illegal" successive onset detection. In the ***RDF*** model, the dynamics of the

resonating process already limit the risk of detecting events close to one another. Therefore, an explicit additional refractory period is unnecessary. We note that, nevertheless, we still consider a short refractory period (35 ms), but its only purpose is to avoid technical issues raised and discussed in the previous chapter, regarding the incapacity of having a full working pipeline of MFCC feature extraction and random forest model prediction when syllable duration is less than 35 ms.

In practice, the ***RDF*** model outputs time instants where it detects syllable events. This results in a binary output vector over time instants between the start of the audio input to the end: for each time-frame, either an event is detected by the ***RDF*** model (1), or not (0). In other words, the ***RDF*** model outputs an all-or-nothing deterministic vector; to embed this into our framework, we need to convert this deterministic output into a probability distribution, of the form $P(OBU^t \mid \Delta L^t)$, that is to say, the probability that there is an event at time instant t given the envelope signal ΔL^t.

Of course, we require this probability distribution to provide high probability values around syllable events detected by the ***RDF*** model. Therefore, we define $P(OBU^t \mid \Delta L^t)$ as a Gaussian mixture distribution, with as many Gaussian kernels as there are syllable events detected by the ***RDF*** model. The means of the kernels are equal to detected time instants, and we set at 1 ms the variance value σ^2_{BU}, for all the following simulations. Therefore, we tend towards a Gaussian mixture of very narrow kernels, that is to say, with each kernel being "quasi-Dirac", leading to binary, almost "all-or-nothing" responses.

1.1.2 Adapting the top-down onset prediction

To recall, the idea behind the top-down onset prediction is to have another channel of information coming from "higher-level" linguistic knowledge, such as lexical, prosodic, or semantic representations. In Chapter 3, the top-down temporal prediction was lexical and derived from word composition. It relied on typical durations of constituent syllables, supposed to be known for every word in the lexicon, and used to create a Gaussian mixture distribution for each word, with a Gaussian kernel positioned at each expected syllable boundary. In the COSMO-Onset model variant for real speech input, we must replace this lexical knowledge with some other information on the likely distribution of syllable durations. In the following paragraphs, we present three top-down models, of increasing levels of adaptation to the structure of the speech input.

Basic top-down prediction of standard syllable duration The first and simplest model we implement assumes that typical syllable duration is within a preferred range of rhythmicity around 4–8 Hz, in agreement with several studies

on the distribution of syllabic rhythm in languages of the world (Ding et al., 2017; Greenberg et al., 2003; Poeppel & Assaneo, 2020; Varnet et al., 2017).

Speech rate adapted top-down temporal prediction of standard syllable duration The previous "overall syllable duration" model is easily refined by computing a more precise mean syllable duration from a given set of observations. This corresponds to adapting the top-down estimation of syllable duration to the observed speech rate, from the interlocutor, or, in our experimental case, from a reference corpus.

Prosodic top-down temporal prediction of syllable duration The third model we implemented incorporates straightforward prosodic information exploiting the position of a syllable within a sentence. Indeed, it is known that syllable duration differs along sentence production, with, for instance, syllable lengthening in final positions (Ferreira, 1993; Lindblom, 1968). To implement this in our model, we introduce a probability distribution of the form $P(DSyl_{i+1} \mid pos_i)$, enabling us to predict the next syllable onset knowing the current syllable position. This can be computed from the corpus of syllables created from the corpus of audio sentences: we define a Gaussian distribution for each position, with its mean and variance the empirical mean and variance measured in the corpus. Importantly, this model is independent of word identity within the sentence, consequently making it suitable for our syllable type recognition at hand. Therefore, with this probability distribution identified from the corpus, it is easily used in the model: given the known current position p_i in the sentence processing, we select the single Gaussian distribution $P(DSyl_{i+1} \mid [pos_i = p_i])$, to predict the next expected syllabic event.

1.2 Adapting the decoding module

In the general COSMO-Onset model architecture presented in Chapter 2, the decoding module is hierarchically organized with alternating layers about syllables and words. The processes taking place to go from the speech signal to syllable categories were wrapped within a portion of the model we noted **X**. The version of the model that we define here is only concerned with the task of syllable type recognition so that the word layer can be simplified from the general architecture. This implies a slightly modified architecture for the top-down temporal prediction, which, contrary to Figure 2.1, is not informed by a word-level variable, but by syllable-level variables instead (and, in the case of our prosodic top-down model, a variable representing syllable position in the sentence).

To recall, in Chapter 3, syllable recognition relied on an intermediate stage of phone decoding. In the previous Chapter 5, we have trained a random forest

(RF) model to perform syllable type recognition from the raw speech signal by first, extracting the MFCC features and then predicting the syllable type.

The RF model provides, as its output, a probability distribution over the 6 syllable types. Importantly, in this implementation, the online (continuous) decoding process operates at the "syllable level", that is, we update all the states of the model after every syllable onset detected by the temporal control module, and all signals located between two successive syllable onsets are passed onto the RF model to perform syllable type recognition. The **X** portion of the model, linking the input to the syllable layer and expressed in terms of the form $P(SyS^t \mid I^t)$, embeds the decoding mechanisms performed by the RF model.

Altogether, this results in a global architecture for the present COSMO-Onset architecture displayed on Figure 6.1.

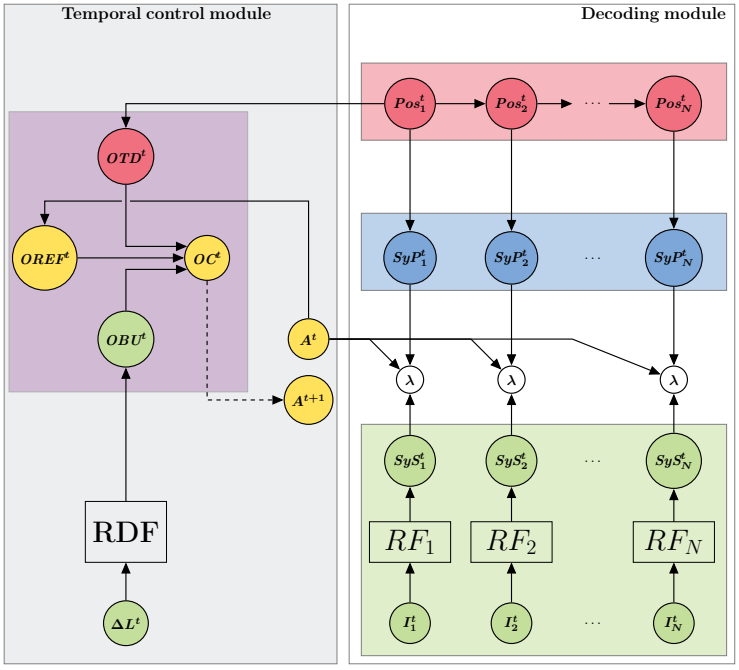

Figure 6.1: Graphical representation of COSMO-Onset for real speech. The legends respect the same taxonomy as on Figure 3.1.

2 Theoretical Hypotheses

Now that we have fully defined the COSMO-Onset model in this new variant, by including and adapting the **X** and **Y** portions, and refining the rest of the model's architecture, we can state the theoretical hypotheses that we are interested in, and that the model simulations will aim at addressing.

The fundamental hypothesis concerns the potential role of top-down information to improve speech perception. This can be evaluated at two levels. First, at the event detection level (**HYPOTHESIS 1**): does the top-down temporal prediction model that we added to the more classical bottom-up onset detection module improve event detection? Second, at the sequence recognition level (**HYPOTHESIS 2**): does it also improve sequence recognition?

In the second stage, we also evaluate whether the complete temporal control module associating a bottom-up oscillator-based onset detection process and a top-down syllabic duration prediction process provides simulations in line with observed behavior. We focus on the experimental data presented by Aubanel and Schwartz (2020), regarding the complementary roles of naturalness and isochrony in speech perception. This leads us to two additional hypotheses. First, we propose that resonance phenomena in the bottom-up ***RDF*** onset detection process would favor isochrony (**HYPOTHESIS 3**): are natural sentences that are more isochronous better recognized than those that are less isochronous? Finally, we propose that "linguistic" predictions in the top-down component of onset detection would favor naturalness (**HYPOTHESIS 4**): are isochronous sentences that are more natural better recognized than those that are less natural?

3 Simulation Material

3.1 Corpus

All the simulations that are performed in this Chapter still exploit the *Fharvard* corpus, as the ones in the previous chapters. Since we are interested in the behavior of the model for sequences of syllables, therefore, instead of segmenting the sentences of the corpus into sets of independent syllables, as we did to train our Random Forest model, we consider here the whole sentences of the corpus.

For our simulations on natural sentences, the corpus was already described (see Section 1.1 of Chapter 4). However, the corpus also contains the same 177 sentences, but in a version rendered isochronous with respect to the constituent syllable P-centers (see Chapter 1, Section 6) [1]. As their natural versions, these 177 isochronous sentences are fully annotated at various levels, including phonemic, syllabic, and word levels. These isochronous versions of the sentences will be used to assess our fourth Hypothesis.

The sentences of the corpus can be characterized by two measures. The first is the distortion metric characterizing the departure of natural sentences from isochrony (*diso*), which we previously used in Chapter 4. The second is the distortion metric characterizing the departure of isochronous sentences to

[1] Examples of stimuli can be found here, supplementary information 2.

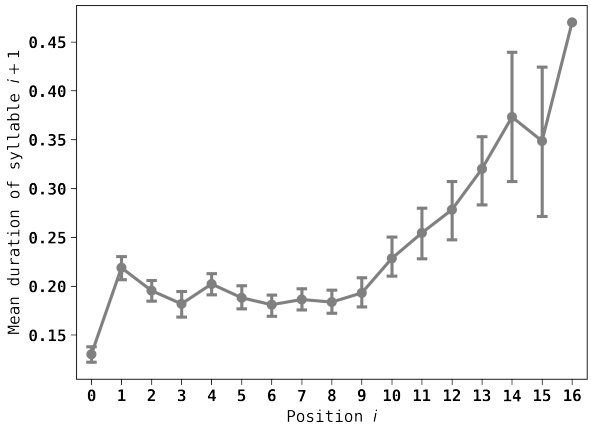

Figure 6.2: Syllable duration statistics for the prosodic top-down temporal prediction model. The plot shows the mean and standard deviation (error bars) of syllable duration in the *Fharvard* corpus (y-axis) as a function of syllable position in the sentence (x-axis). As reasonably expected, uncertainty and hence variance increase with position and reach rather large values towards the last positions in the sentence. A single sentence contains 17 syllables (which is the maximum value in the corpus), we, therefore, do not show the standard deviation for the right-most data point.

naturalness (*dnat*).

3.2 The decoding module

We use the random forest model built in the previous chapter for syllable recognition. We thus keep this portion of the model exactly as it was defined, with all its parameters and optimal hyper-parameters.

3.3 The bottom-up model of the temporal control module

We use the ***RDF*** model presented extensively and evaluated in Chapter 4 in its original version designed to detect syllabic onsets. It is also used exactly as it was defined, with its optimal parameters.

3.4 The variants of the top-down model of the temporal control module

We consider three variants for the top-down model of the temporal control module. The first one can be considered as the *baseline* top-down temporal model, with an assumed speech rate at 8 Hz, and therefore a mean syllable duration of $1/8 = 125\ ms$. This can be viewed as a sort of non-informed preferred rate

(intrinsic rhythm). We will refer to this top-down model variant as the **top-down with flat speech rate**.

For the second variant, we computed the mean speech rate from the *Fharvard* corpus. This informed preferred rate provides a mean syllable duration equal to 150 ms and a variance equal to 60 ms. We will refer to this top-down model variant as the **top-down with mean speech rate**.

The third and final variant relies on prosodic knowledge, predicting the next onset based on the current syllable position. Using the *Fharvard* corpus for the 177 sentences, we computed the statistics of syllable duration (mean and variance duration) as a function of syllable position; this is displayed on Figure 6.2 and shown in Table 6.1. We will refer to this third variant as the **top-down with position**.

3.5 Performance measures

Depending on the hypothesis (one of the four aforementioned), we either use the event detection metrics (boundary performance) or the unit identity metrics (unit performance), or the combination of both (temporal overlap) presented in Section 5.1 of Chapter 1.

In addition to the set of natural or isochronous speech stimuli, we also tested the model on noisy stimuli to assess the model's performance in degraded conditions. We hence added white noise at various intensities and we considered 3 levels of signal-to-noise ratio: -10 dB, -20 dB, and -30 dB.

The four sets of natural or isochronous stimuli, without noise or degraded at 3 SNR levels, were processed with four variants of the temporal control model: a bottom-up only variant, and the whole model with the three variants of top-down prediction models. For all variants with top-down predictions, we only experimented with the *AND* fusion operator, which seems better adapted to possibly filter out spurious events likely to emerge from the bottom-up processing branch, particularly in noise.

Whether it is the bottom-up only model variant or the full model with the fusion of bottom-up onset detection and top-down temporal prediction, we compute "continuously", at each time step (assumed to correspond to 1 ms), the probability that there is a syllabic onset, as evaluated from the temporal control module. Whenever this probability value reaches a threshold, that is set to 0.3 for all the following simulations, the temporal control module decides that a syllable boundary is reached, which results in advancing in the syllable processing sequencing (see Table 6.1 for the summary of the temporal control module). This has several effects in the model; to recall, the current syllable is then considered terminated, the next one begins to be processed, and the position counter pos_i is incremented.

Table 6.1: Summary of the temporal control module parameters. The first column lists model variants, and, for the top-down with position model, positions. The second column contains Gaussian kernel means expressed in ms. The third column contains Gaussian kernel standard deviations expressed in ms.

Model		Mean (ms)	Standard deviation (ms)
Bottom-up only		Correspond to detected events by the ***RDF*** model	1
TD with flat speech rate		125 ms after each detected onset	60
TD with mean speech rate		150 ms after each detected onset	60
TD with position	0	130	54.9
	1	218	73.6
	2	195	73.5
	3	182	86.7
	4	202	74.6
	5	188	78.1
	6	181	74.2
	7	186	75.2
	8	184	78.8
	9	193	101
	10	228	131
	11	254	147
	12	278	149
	13	320	138
	14	373	162
	15	348	117
	16	470	50

4 Simulation Results

4.1 Illustrative example of the whole model

Figure 6.3 provides an illustrative example of applying the model on a complete sentence. The bottom plot of Figure 6.3 shows the representation of the input corresponding to the sentence "La lampe de néon rouge irise ses cheveux" (in English: The red neon lamp makes her hair glow) which lasts about 2.75 s. It shows the energy amplitude, the normalized amplitude envelope and the ground truth syllable boundaries. The syllabic annotation for the sentence is the following sequence (from the annotated data by Aubanel and Schwartz (2020)): *"la lã pə də ne õ ʁu ʒə i ʁi zə se ʃə vø"*. Converting this syllable sequence into syllable types yields: *CV, CV, CV, CV, CV, VC*, CV, CV, VC*, CV, CV, CV, CV, CV*. We notice that this example mostly contains CV syllables, which account for nearly 86 % (12 / 14) of the sentence.

The rest of the plots (all other plots except the last bottom of the left part) on Figure 6.3 shows the simulation results of the different components of the temporal control module of the model. The bottom-up onset detection of the ***RDF*** model (Figure 6.3, the second bottom left plot) detects 18 events (adding the first onset of signal). The top-down onset model using the prosodically-informed variant, predicting syllable duration from their position (Figure 6.3, top left plot), almost

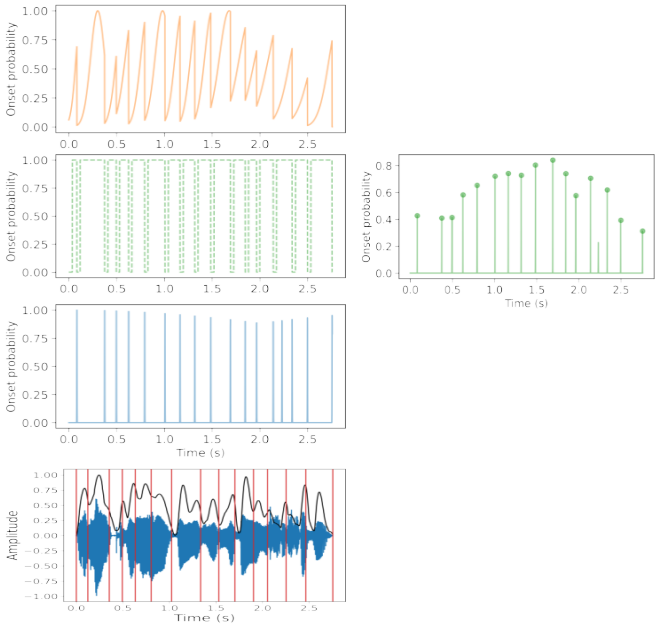

Figure 6.3: Illustrative example of applying the model on a complete sentence. Bottom plot: Representation of the energy amplitude (in blue) and the normalized amplitude envelope (in black), and the ground truth syllable boundaries (in red vertical lines) from the annotated corpus *Fharvard* by Aubanel and Schwartz (2020) for the sentence "La lampe de néon rouge irise ses cheveux" (in English: The red neon lamp makes her hair glow) as a function of time (x-axis). The four plots above the bottom plot illustrate the evolution of onset probability (y-axis) in the temporal control module, as a function of time (x-axis). Left column, from top to bottom: in orange, the top-down onset prediction, in dashed green, the probability of an onset being outside a refractory period, in blue, the bottom-up onset detection by the ***RDF*** model. Right plot: in green, the output of the temporal control module for evaluating onset probability, according to the AND fusion model. Note that, for all plots except the last bottom plot, the first initial onset is not represented, although it is taken into consideration in all simulations for performance measures.

predicts the same events as the bottom-up onset detection model, except for one onset located near 2.23 s. These two sets of onsets combined with the refractory onsets (Figure 6.3, the second top plot) with the *AND* fusion operator result in the final onsets that would be considered the outputs of the temporal control module (Figure 6.3, right plot). We observe that, in this example, the onset detected by the bottom-up component at 2.3 s is filtered out by the top-down prediction. Comparison with the ground truth annotation (red bars on the bottom left plot of Figure 6.3) shows that this removal is, in this case, an error.

At each detected onset (the first onset being the one following the signal onset and the last one corresponding to the signal offset), the signal between the last onset and the previous one is sent to the decoding module (the random

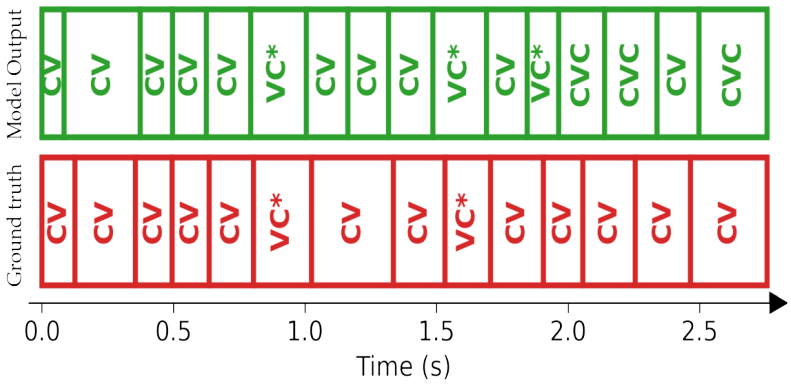

Figure 6.4: Results of the decoding module recognizing the syllable types for the sentence "La lampe de néon rouge irise ses cheveux" (in English: The red neon lamp makes her hair glow). Syllable types in red correspond to ground truth syllable types annotated within their corresponding temporal intervals (red boxes). Syllable types in green correspond to recognized syllable types by the random forest model, also within their corresponding intervals (green boxes).

forest model) to perform syllable type recognition. Figure 6.4 shows the results of the decoding module on this illustrative sentence. Overall, we observe that the temporal control module detects onset events at relevant instants, thus providing segments of acoustic input to the decoding module that are rather well-bounded, resulting in relatively few errors both for syllabic onset detection and for syllable type recognition. In this illustrative example, concerning event detection, model performance evaluation yields a recall of 80 %, a precision of 70.6 %, and an F-score of 75 %. To recall, the event detection performance scores are measured by allowing a margin of error of 50 ms (before and after real onsets). Here, since the model missed three onsets located respectively at 1.905 s, 2.055 s, and 2.255 s, and added five onsets at 1.16 s, 1.845 s, 1.964 s, 2.139 s, and 2.337 s, the model has a recall of $12/15 = 0.8$ and precision of $12/17 \approx 0.706$, which leads to an F-score of 75 %.

Comparing the sequence of syllable type output by the model on Figure 6.4, we observe some errors. Using the sequence matching computation based on the Levenshtein distance (Young et al., 2002; Yujian & Bo, 2007), we get a number of deletions of 2 (the predicted syllable types at positions 7 and 16) and a number of substitution of 3 (predicted syllable types at positions 12, 13 and 14). This gives us a percent accuracy of 64.3 %. However, based on the temporal overlap of predictions and the ground truth, the model score is at 70 %. Indeed, the temporal overlap measure penalizes fewer insertions and deletions compared to the Levenshtein distance score. Consider for instance the time interval between 1.0 and 1.4 s: a single long CV syllable is the expected answer, and the model is

Table 6.2: The role of top-down in syllabic event detection for natural stimuli. The first column describes the experimental condition (without noise, or with added noise at varying SNR: -10 dB, -20 dB, -30 dB). The second column gives the model variant: in blue the bottom-up only model, in teal the full model combining bottom-up and top-down with flat speech rate, in purple the full model combining bottom-up and top-down with mean speech rate and in orange the full model combining bottom-up and top-down with the syllable position. In the following columns (3, 4, 5), the syllabic event detection measures are expressed in percentage, respectively the F-score, the precision, and the recall. For every condition and for every measure, the best score is displayed in bold.

Condition	Model variant	F-score	Precision	Recall
No noise	BU-Only	77.93	71.26	**86.51**
	TD with flat speech rate	74.76	71.53	79.09
	TD with mean speech rate	76.46	72.79	81.28
	TD with position	**78.30**	**73.50**	84.19
−10 *dB* SNR	BU-Only	61.77	46.31	**93.58**
	TD with flat speech rate	61.64	46.47	92.42
	TD with mean speech rate	65.17	51.73	88.67
	TD with position	**66.91**	**58.28**	78.99
−20 *dB* SNR	BU-Only	55.28	39.72	**91.59**
	TD with flat speech rate	55.51	40.03	91.25
	TD with mean speech rate	**60.79**	46.67	87.80
	TD with position	60.21	**51.9**	72.07
−30 *dB* SNR	BU-Only	55.11	39.73	**90.66**
	TD with flat speech rate	55.52	40.18	90.55
	TD with mean speech rate	**60.44**	46.61	86.61
	TD with position	59.65	**51.36**	71.52

correct on syllable type but considers it as two syllables, erroneously introducing a syllable frontier around 1.2 s. This impacts the Levenshtein distance score (an insertion), but not the temporal overlap score (at each instant in this interval, a CV syllable type is correct).

4.2 Contribution of top-down predictions in syllabic event detection

We now return to the analysis of our simulation on the whole corpus. Table 6.2 shows the experimental performance measures for syllable onset detection for the different model variants in the different experimental conditions. Figure 6.5 reprises the experimental F-score measures (third column of Table 6.2) in graphical form.

In the baseline experimental condition, that is, without noise, the F-score performance indicates that all models have similar performance overall. Interestingly, the "bottom-up only" variant of the model is always better in recall than the other model variants combining bottom-up with top-down. However, it is less precise compared to the other models. In degraded conditions, we first observe an overall performance decrease for all variants, as noise increases. Finally, we observe a slight performance increase for top-down variants compared to the bottom-up only model, with the top-down with position variant being the best or very close to the best model.

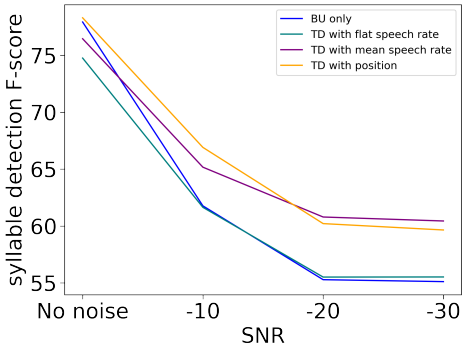

Figure 6.5: Experimental F-scores in syllabic event detection for natural stimuli. The plot shows the F-score of experimental results (y-axis) as a function of signal-to-noise ratio (x-axis), for all 4 model variants: bottom-up only in blue, the full model with the top-down prediction with flat speech rate in teal, the full model with the top-down prediction with mean speech rate in purple and full model with the top-down prediction with syllable positions in orange. Data shown here in graphical form is identical to the F-scores reported in column 3 of Table 6.2.

Considering only the different top-down model variants, we observe that the worst-performing model at event detection is the one with a non-informative speech rate. Interestingly, the more informed the model variant, the higher its precision, and the least informed the model variant, the higher its recall. Considering the F-score measure, which is an aggregate of precision and recall, we observe that the two best model variants are the "TD with mean speech rate" and the "TD with position" variants, especially in conditions with a high level of noise.

4.3 Contribution of top-down predictions in syllabic sequence recognition

Table 6.3 shows the summary of the syllabic sequence recognition performance of all model variants in the different experimental conditions. Figure 6.6 represents the experimental temporal overlap measures (right column of Table 6.3) in graphical form.

We recall that the accuracy measure compares the sequence of syllable types evaluated by the model with the ground truth sequence, and is a measure that is rather sensitive to insertions and deletions. The second measure, the temporal overlap metric, combines both syllable type recognition and boundary detection. It counts the portion of time where syllable type matches between the model output and ground truth. The pattern of performances is quite clear, with an increase in performance from the BU-Only to the TD with flat speech rate to the TD with mean speech rate, and finally to the TD with position variant. Differences are very large in the accuracy criterion and particularly in noise.

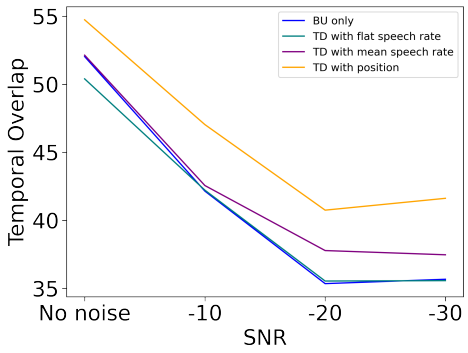

Figure 6.6: Experimental temporal overlap in syllabic sequence recognition for natural stimuli. The plot shows the temporal overlap scores (y-axis) as a function of signal-to-noise ratio (x-axis), for the bottom-up only variant in blue, the full model with the top-down prediction based on flat speech rate in teal, the full model with the top-down prediction based on mean speech rate in purple, and the full model with the top-down prediction with syllable positions in orange. Data shown here in graphical form is identical to the temporal overlap scores reported in the right-most column of Table 6.3.

Table 6.3: The role of top-down in syllabic sequence recognition for the natural stimuli conditions. The first column describes the experimental condition (without noise, or with added noise at varying SNR: -10 dB, -20 dB, -30 dB). The second column gives the model variant with 4 possibilities. In the following columns (3, 4), the syllabic sequence recognition measures are expressed in percentage, respectively the accuracy and the temporal overlap. For every condition and for every measure, the best score is displayed in bold.

Condition	Model variant	Percentage Accuracy	Temporal Overlap
No noise	BU-Only	48.64	52.04
	TD with flat speech rate	52.04	50.40
	TD with mean speech rate	52.47	52.14
	TD with position	**53.54**	**54.74**
-10 *dB* SNR	BU-Only	1.18	42.15
	TD with flat speech rate	1.57	42.22
	TD with mean speech rate	9.30	42.56
	TD with position	**34.73**	**47.04**
-20 *dB* SNR	BU-Only	0	35.35
	TD with flat speech rate	0.04	35.54
	TD with mean speech rate	2.77	37.78
	TD with position	**26.76**	**40.75**
-30 *dB* SNR 0.27183170214077496, 0.416216803	BU-Only	0.084	35.67
	TD with flat speech rate	0.084	35.57
	TD with mean speech rate	3.14	37.47
	TD with position	**27.18**	**41.62**

But the pattern is also clear with the temporal overlap measure. Altogether it appears that the temporal prediction model based on syllable position within the sentence is the best one and provides a strong improvement in performance. In the following, it will be the only TD model that we keep in comparison with the BU variant.

Similar to the natural corpus evaluation, in Table 6.4, we show model performance on the corpus of sentences rendered isochronous. We observe the same performance tendencies and comparison patterns between the model variants. Somewhat unexpectedly, we observe an overall slightly better performance on

Table 6.4: The role of top-down in syllabic sequence recognition for isochronous stimuli conditions. The table format is identical to Table 6.3.

Condition	Model variant	Percentage Accuracy	Temporal Overlap
No noise	BU-Only	53.85	55.28
	TD with position	**55.30**	**55.75**
−10 *dB* SNR	BU-Only	2.16	44.37
	TD with position	**33.7**	**48.99**
−20 *dB* SNR	BU-Only	0	37.96
	TD with position	**27.84**	**43.61**
−30 *dB* SNR	BU-Only	0.16	37.73
	TD with position	**27.71**	**42.64**

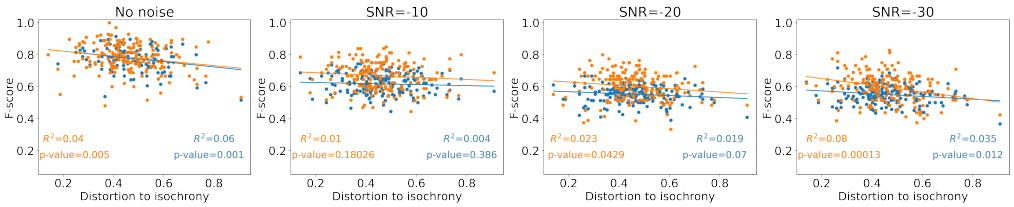

Figure 6.7: Role of isochrony in syllabic onset detection for natural sentences. All plots (from left to right: no noise, SNR at −10 *dB*, −20 *dB* and −30 *dB*) show the experimental F-score (y-axis) as a function of distortion to isochrony (x-axis), both for the bottom-up only model variant (in blue) and the full model combining the bottom-up detection and the top-down with position (in orange). Linear regressions (solid lines) and corresponding squared correlation coefficients R^2 and p-values are also indicated in the plots.

isochronous stimuli compared to natural stimuli.

4.4 Role of isochrony in speech perception for natural sentences

Figure 6.7 shows event detection performance as a function of distortion to P-center isochrony, for syllable boundary detection in the different experimental SNR conditions, on the naturally timed corpus. We observe that there is indeed a statistically significant negative correlation between model performance and temporal distortion in most cases, which confirms that model performance improves for more isochronous sequences (i.e., with lower temporal distortion to isochrony). Interestingly, the effect appears already in the bottom-up model variant. Nevertheless, the top-down variant does not remove this property and possibly amplifies it, with larger correlation values (compare the orange with the blue values in the figure). There is also an overall slight gain in performance with the model including top-down predictions.

Figure 6.8 shows the experimental measure of syllabic sequence recognition, as measured by percent accuracy, as a function of temporal distortion from isochrony. We observe a large difference in performance between the "BU-only" and "TD

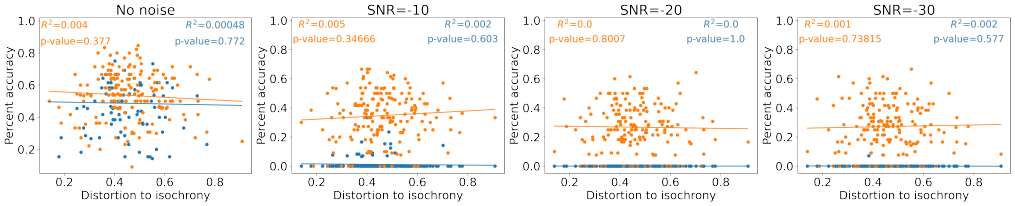

Figure 6.8: Role of isochrony in syllabic sequence recognition for natural sentences. Plots show experimental accuracy in sequence recognition (y-axis) as a function of distortion to isochrony (x-axis); the rest of the graphical representation and organization is identical to Figure 6.7.

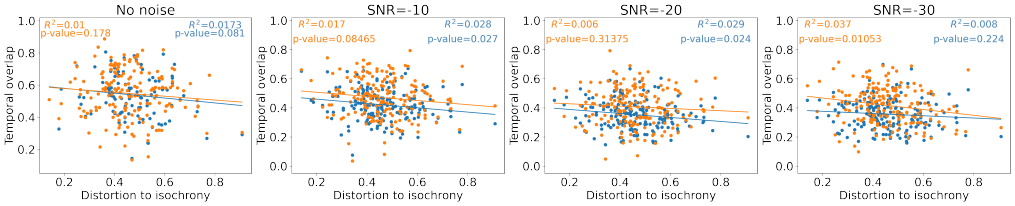

Figure 6.9: Role of isochrony in the temporal overlap measure for natural sentences. Plots show experimental temporal overlap in sequence recognition (y-axis) as a function of distortion to isochrony (x-axis); the rest of the graphical representation and organization is identical to Figure 6.7.

with position" variants, especially in noise (already observed from aggregated data of Table 6.3. However, isochrony does not appear to play any significant role here, since regardless of the experimental condition, that is, whether there is noise or not, we do not find a statistically significant negative correlation between model performance and distortion to isochrony.

Finally, we display on Figure 6.9 the experimental results for temporal overlap. In all conditions, we observe both the (sometimes small) gain associated with top-down predictions and the benefit of isochrony with significant negative correlations between distortion to isochrony and performance in most conditions. Once again, the effect of isochrony already appears with the "bottom-up only" variant and is only slightly affected by the addition of top-down predictions.

4.5 Role of naturalness in speech perception for isochronous sentences

Figure 6.10 shows event detection performance as a function of distortion to the temporal distribution of natural P-centers ("distortion to naturalness"), for syllable boundary detection on the corpus of sentences rendered isochronous, in the different noise conditions. We observe in most cases a slight but statistically significant negative correlation between model performance and temporal distortion, meaning that model performance is higher when the temporal distortion from a natural distribution of events is small. Therefore, for isochronous sentences

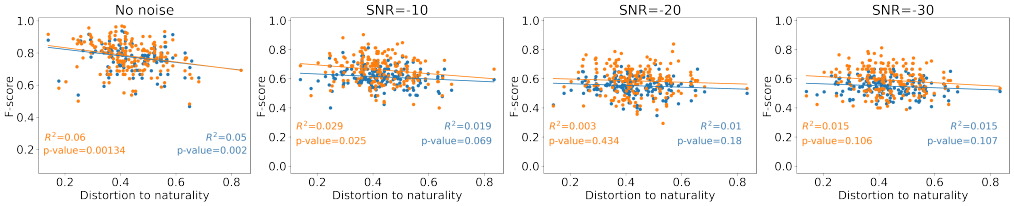

Figure 6.10: Role of naturalness in syllabic onset detection for isochronous sentences. Plots show experimental F-scores in sequence detection (y-axis) as a function of distortion to naturalness (x-axis); the rest of the graphical representation and organization is identical to Figure 6.7.

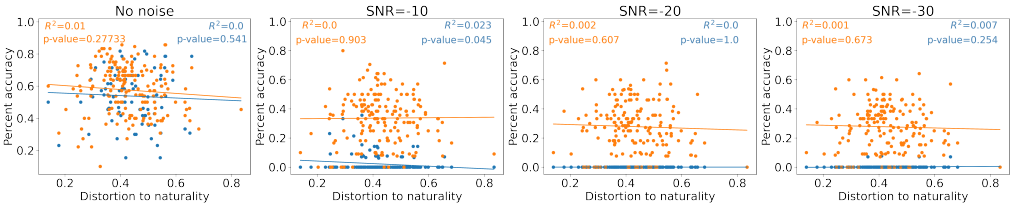

Figure 6.11: Role of naturalness in pure syllabic sequence recognition for isochronous sentences. Plots show experimental accuracy in sequence recognition (y-axis) as a function of distortion to naturalness (x-axis); the rest of the graphical representation and organization is identical to Figure 6.7.

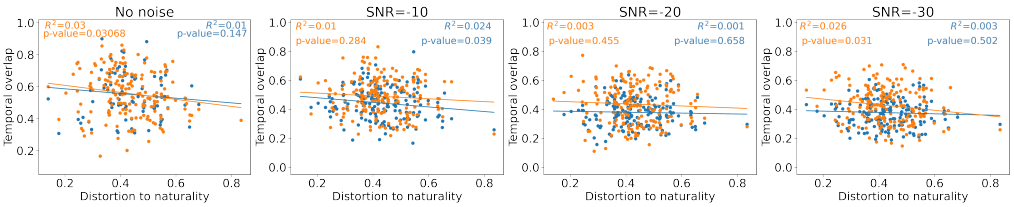

Figure 6.12: Role of naturalness in the temporal overlap performance for isochronous sentences. Plots show experimental temporal overlap in sequence recognition (y-axis) as a function of distortion to naturalness (x-axis); the rest of the graphical representation and organization is identical to Figure 6.7.

which happen to be closer to naturally-timed sentences, model performance is higher. Once again, we observe a slight top-down effect with better performance when top-down predictions are added. Nevertheless, it appears that correlations between event detection performance and naturalness are already found for the "bottom-up only" model.

Concerning syllabic sequence recognition, accuracy scores displayed on Figure 6.11 show almost no effect of naturalness on model performance. However, we still observe a large effect of the top-down branch, which strongly improves accuracy in all conditions.

Finally, temporal overlap scores (Figure 6.12) show a slight gain in performance with the top-down branch and a significant negative correlation between distortion to naturalness and performance in various cases. Once again, and still, surprisingly,

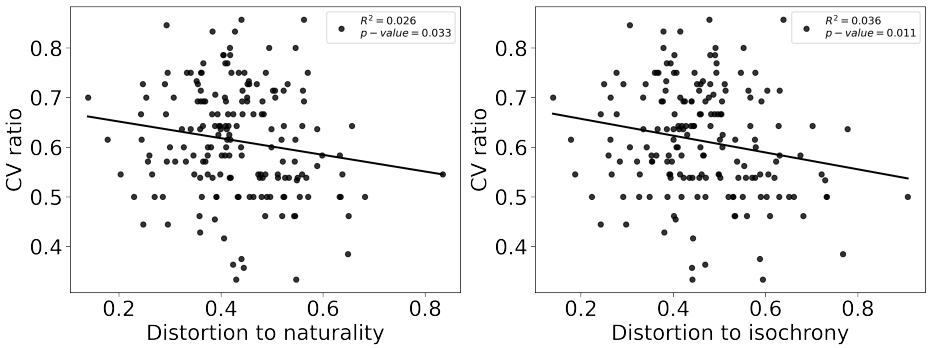

Figure 6.13: Relation between the CV ratios and the distortion values. On the left: plot of the CV ratio on the y-axis vs the distortion to naturalness on the x-axis for isochronous sentences. On the right: plot of the CV ratio on the y-axis vs the distortion to isochrony on the x-axis for natural sentences. On both plots, the linear regression coefficients (R^2) and the interval of confidence are also displayed

negative correlations are also found for the "bottom-up only" variant.

5 Discussion

5.1 Summary of main simulation results in relation to the four hypotheses

In this chapter, we combined our model pieces together, in order to define a variant of the COSMO-Onset model able to deal with real speech input. We have evaluated model performance on the *Fharvard* French corpus, using both sentences with their original temporal organization, and sentences rendered isochronous. This corpus was used previously by Aubanel and Schwartz (2020), which allows comparison. Globally, it appears that the four tested hypotheses are confirmed in the present simulation.

For the first hypothesis, we focused on the effect of top-down temporal prediction in syllabic onset detection. We tested this hypothesis on naturally timed sentences, without noise, and in noisy conditions. First, we observe that in all conditions, the bottom-up only model is the best model with regard to the recall metric. This is possibly due to the fact that the ***RDF*** model is rather sensitive, with a tendency to detect more events than there really are in the signal. When it detects events that are "spurious", this is not reflected in the recall scores, but it is in the precision scores. Indeed, we observed good recall for the BU-only variant. Top-down models, on the other hand, filter out spurious events detected by the bottom-up only model, leading to increasing precision and decreasing recall. F-scores, which combine precision and recall, yield a small advantage to models with top-down knowledge. Second, in the baseline experimental condition

(stimuli without noise), all model variants have similar performance measures. This is consistent with the fact that in the case of no difficulty, i.e., when the natural signal is not degraded, the contribution of the top-down is very small or null, probably because the information from the bottom-up and top-down systems are redundant. However, generally, we observed a performance increase, that is to say, a positive effect of the top-down prediction for syllabic onset detection, with the prosody-based top-down model being globally (in scores averaged over the four noise conditions) the best among the four model variants tested (bottom-up only, top-down with uninformed speech rate, top-down with informed speech rate and the prosodically-informed top-down model). Still, at this stage, the difference between the last two variants is quite small.

For the second hypothesis, we studied the contribution of top-down predictions to syllable sequence recognition on the naturally timed and isochronous sentences, in all noise conditions. Again, we found a positive effect of top-down predictions on performance, with a larger performance increase. Crucially, the prosodically informed variant providing temporal information in relation to the position of the syllable in the sentence is by far the best one. This is due to the fact that this model provides a significant increase in the precision of event detection, which appears to automatically result in the elimination of many spurious events, thus increasing recognition scores.

Concerning the third and fourth hypotheses, they are also both confirmed in the sense that we obtained significant regressions with negative slopes, both for natural sentences between departure from isochrony and event detection F-scores (Figure 6.7) or temporal overlap in syllable recognition (Figure 6.9) (third hypothesis), and for isochronous sentences between departure from naturalness and event detection F-scores (Figure 6.10) or temporal overlap in syllable recognition (Figure 6.12) (fourth hypothesis). Of course, it is important to stress that, though regressions are significant and often highly significant in a number of plots within the corresponding figures just mentioned, they actually explain a small part of the total variance in these plots (just a few percent). This is actually not surprising concerning the large variability of the presented acoustic and phonetic material, and this is similar to what was found in the experimental data in Aubanel and Schwartz (2020) (see their Tables 2, 3, and 4). Still, the trends are rather systematic, confirming that in COSMO-Onset also, "natural and isochronous are both beautiful", as in the experimental data in Aubanel and Schwartz (2020).

5.2 Two intriguing results provided by the simulations

Still, two intriguing results emerged from our simulations. Firstly, it appears that isochronous sentences are better recognized than natural ones. Indeed, temporal

Table 6.5: The role of top-down in syllabic event detection for isochronous stimuli. The table format is identical to Table 6.2.

Condition	Model variant	F-score	Precision	Recall
No noise	BU-Only	77.56	72.50	**83.78**
	TD with flat speech rate	73.04	72.56	74.21
	TD with mean speech rate	75.04	72.34	78.69
	TD with position	**78.16**	**73.5**	82.96
−10 *dB* SNR	BU-Only	61.32	46.26	**91.9**
	TD with flat speech rate	61.4	46.52	91.25
	TD with mean speech rate	64.42	51.21	87.59
	TD with position	**66.06**	**57.6**	77.91
−20 *dB* SNR	BU-Only	55.24	39.67	**91.74**
	TD with flat speech rate	55.59	40.06	91.58
	TD with mean speech rate	**60.15**	45.99	87.48
	TD with position	58.57	**50.54**	70.03
−30 *dB* SNR	BU-Only	54.97	39.61	**90.65**
	TD with flat speech rate	55.27	39.96	90.46
	TD with mean speech rate	**60.32**	46.46	86.65
	TD with position	59.06	**50.91**	70.74

overlap scores in Tables 6.4 vs. 6.3 are 1.5 to 3 % better in the first case. This is not the case at the level of event detection (see Table 6.5). Importantly, this result is at odds with the experimental data, which show a clear decrease in recognition accuracy for isochronous compared with natural stimuli (see Fig. 1.9, from the paper by Aubanel and Schwartz (2020)). The reason in our view could be the fact that isochronous sentences are produced by shortening long inter P-center intervals and lengthening short intervals, which probably results in shortening long syllables and lengthening short ones. Indeed, the RF model presented in Chapter 5 displays much better recognition scores for CV syllables than for all other ones, and CV syllables are probably typically shorter than other syllable types (e.g. CVC, CCV, CCVC, etc.). This likely increases somewhat artificially temporal overlap scores for isochronous material.

The second puzzling fact concerns the behavior of the bottom-up ***RDF*** model in these simulations. As a matter of fact, coming back to the "natural and isochronous are both beautiful" claim, we expected the role of isochrony to be driven by bottom-up resonance processes, and the role of naturalness to be driven by top-down predictions. In Figure 6.7, it appears indeed that the role of isochrony for the detection of onset events is already present in the bottom-up only variant, which confirms that it should be driven by resonance processes in the ***RDF*** model. This is in line with the simulations in Chapter 4. However, Figure 6.10 shows that there is also an effect of naturalness for onset detection in the bottom-up only variant, at least in the condition without noise. This was absolutely not expected. Indeed, in this case, all sentences are isochronous and were expected to display basically no difference, and certainly no difference in relation to their smaller or larger naturalness. There is possibly an increase in regression value when top-down predictions are added, at least without noise or at an SNR level of -10 dB, but the bottom-up effect remains puzzling.

We conjecture that this could be due to the fact that isochronous sentences

that are close to their natural counterpart, that is, sentences that are not distorted much when rendered isochronous, are also those that have the most regular syllabic structure. This would probably make their envelope more regular and hence onset detection easier. To test this hypothesis, we estimated for each sentence the ratio of CV syllables over the total number of syllables in each sentence. We display in Figure 6.13 the relationship between this ratio and the distance to the natural distribution for isochronous sentences, and conversely the relationship between this ratio and the distance to the isochronous distribution for natural sentences. This confirms our conjecture. It suggests that the regularity of the phonetic material favors onset detection in the ***RDF*** algorithm, and intervenes in both the effect of isochrony and naturalness in our results (see Figure 6.14 for two examples of the most and the least isochronous natural sentences, and Figure 6.15 their equivalent for isochronous sentences). This unexpected finding adds to our understanding of the potential role of oscillatory-driven neurally inspired algorithms in the temporal processing of speech material.

Altogether, in this chapter, we showed that, even though the pure bottom-up onset detection mechanism achieves relatively good results in event detection, there is still a significant gain in combining with top-down temporal predictions. If in natural situations without signal degradation, the increase in performance is tiny, in degraded situations, the role played by top-down predictions appears more marked. Importantly, regarding the comparison with the study by Aubanel and Schwartz (2020), we showed that the COSMO-Onset model is able to partially replicate the results related to the role of isochrony in natural sentence decoding, and to the role of naturalness of isochronous sentences.

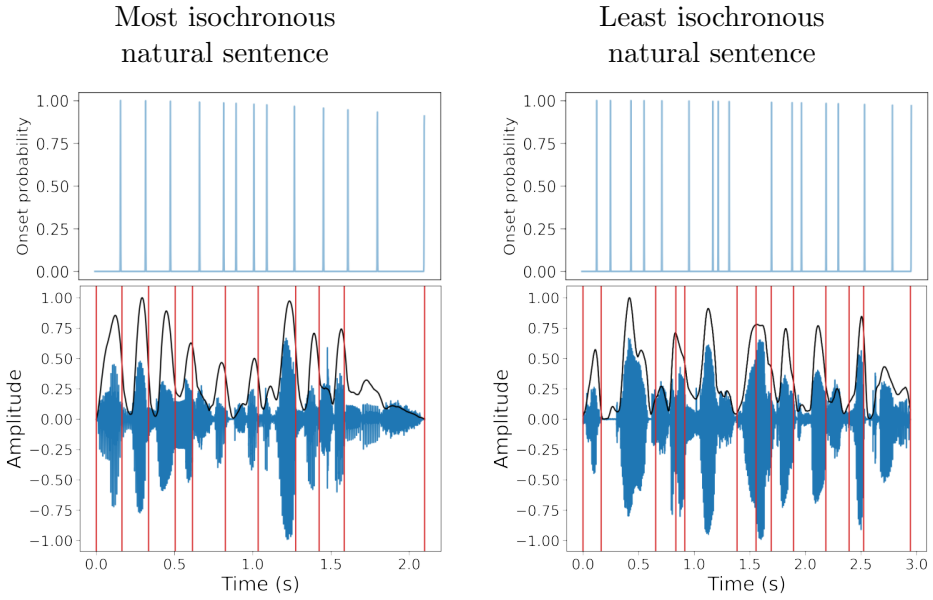

Figure 6.14: Example of sentences at both extremes of the isochrony measure. Left column: the most isochronous natural sentence with a distortion to isochrony value of 0.139 (utterance: "Le bébé met son pied droit dans sa bouche" in French, translated to "The baby puts its right foot in its mouth" in English). Right column: the least isochronous natural sentence with a distortion to isochrony value of 0.907 (utterance: "Les plinthes sur la gauche du hall d'entrée se décollent" in French, translated to "The plinths on the left side of the entrance hall are peeling off" in English). Bottom plots: sentence waveform (blue) with the envelope (black) and ground truth syllable boundaries (vertical red lines). Top plots: bottom-up only model variant boundary detections (in light blue)

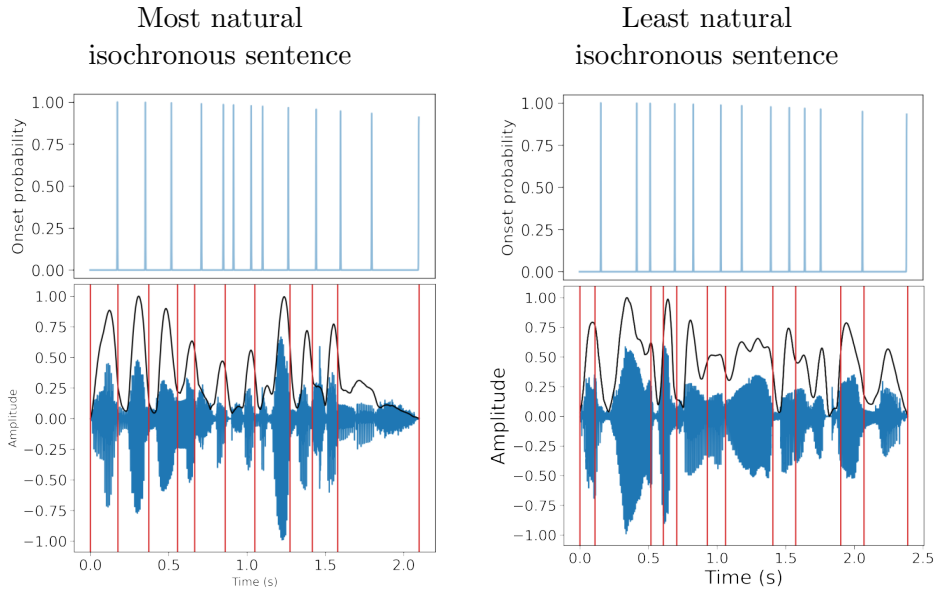

Figure 6.15: Example of sentences at both extremes of the naturalness measure. Left column: the most natural isochronous sentence with a distortion to naturality value of 0.139 (utterance: “Le bébé met son pied droit dans sa bouche” in French, translated to “The baby puts its right foot in its mouth” in English). Right column: the least isochronous natural sentence with a distortion to isochrony value of 0.834 (utterance: “La vieille horloge sur le mur indique midi” in French, translated to “The old clock on the wall indicates noon” in English). Bottom and top plots are arranged in the same order as on Figure 6.14.

Chapter 7

Conclusion and General Discussion

1 Summary of the contributions

In this thesis, we developed and studied COSMO-Onset, a Bayesian model of speech perception with a temporal treatment inspired by neural oscillations. Our main goal was to study the role of top-down temporal predictions in speech segmentation in order to better understand the mechanisms at play in the syllabic segmentation of speech input. We used the COSMO-Onset model to illustrate cases where the top-down information turns out useful in achieving more reliable and accurate segmentation. Especially, even though the role of top-down information in speech perception is widely acknowledged in the literature, the novelty of our work comes from the special care we took to design and develop a fusion model combining both bottom-up onset detection and top-down temporal prediction. The COSMO-Onset model was then compared to findings from a key experimental study by Aubanel and Schwartz (2020) on the role of top-down information in speech perception in noise. We now summarize the main contributions of this thesis.

First, we designed the COSMO-Onset model, a Bayesian speech perception model in line with recent research in speech neuroscience, dissociating the spectro-temporal content decoding, which we called the "decoding module" on the one hand, and the segmentation mechanism, which we called the "temporal control module" on the other hand. This contrasts with classical models of speech perception such as TRACE, which only feature a decoding module. In the temporal control module of COSMO-Onset, there are two main functional parts, the first performing bottom-up onset detection based on the analysis of the envelope of speech input, and the second performing top-down temporal prediction based on higher-level linguistic information. We particularly focused

on **the fusion of bottom-up onset detection mechanisms based on speech envelope processing with neurally inspired oscillatory-based models of speech perception and top-down temporal predictions exploiting lexical or prosodic information about syllable duration**, which to the best of our knowledge is an original feature that current neuro-computational models of speech perception lack.

Second, we developed a first implementation of the COSMO-Onset model to illustrate its core principles on a set of synthetic toy stimuli to perform word recognition and study the interaction of its components. We showed that, in the case of nominal conditions, the use of temporal segmentation cues provided by bottom-up onset detection mechanisms are sufficient to yield high event detection and unit recognition performance. However, in the case of degraded conditions, bottom-up onset detection is less reliable. We showed that properly combining bottom-up cues with top-down temporal predictions allows recovering performance comparable to the nominal condition. We developed two fusion models depending on the degraded condition at hand. The first one is the *AND* fusion model concerned with removing spurious events detected by the bottom-up onset detection mechanisms, for instance, due to noisy signals. The second is the *OR* fusion model concerned with recovering missed events by the bottom-up onset detection mechanisms, for instance, due to hypo-articulation.

Third, we studied the fully oscillatory ***RDF*** model developed by Räsänen et al. (2018). We evaluated its event detection capabilities on the French corpus developed by Aubanel et al. (2020). To do so, **we considered how the model performs with respect to syllable onset detection, which is its original target task, and with respect to P-center detection, for which we extended the original model**. For the former task, the ***RDF*** model performs as well as previous evaluations on other languages (Finnish, Estonian and English) studied by the authors. Furthermore, it appears better to detect P-centers than syllabic onsets, probably because P-centers have more "prominent and robust" markers within the speech envelope dynamics. Importantly, our results showed the role of resonance mechanisms in event detection. We reported two main findings. The first is that **the *RDF* model performs better for resonant than non-resonant features**, and the second is that **speech events within natural sentences with larger inter-P-center isochrony are better detected than those within sentences with smaller inter-P-center isochrony**. This last result is similar to observations from the study by Aubanel and Schwartz (2020).

Finally, we developed a variant of the COSMO-Onset model able to process real speech stimuli. Rather than performing word recognition as in the conceptual model and its first implementation, this version focuses on the recognition of syllable types sequences (the syllabic structure of sentences in terms of consonant

and vowel composition). In this implementation, we used the oscillatory-based model ***RDF*** developed by (Räsänen et al., 2018) to perform bottom-up syllabic event detection by analyzing the speech envelope, and a random forest machine learning algorithm to perform syllable type decoding. We then used three different models of top-down temporal predictions that combine with bottom-up event detection to accomplish the overall syllable event detection, from a simple top-down model based on a "non-informative" speech rate to a prosodically informed one. Our simulation results corroborate well with our previous results on the initial implementation of the COSMO-Onset model. In the nominal condition (speech stimuli without noise), all models perform quite similarly, both in event detection and unit recognition. However, in noisy conditions, the better the model is informed, the better the model's performance, both in event detection and unit recognition. This holds true for both natural and isochronous sentences, with a small performance gain for isochronous sentences. Furthermore, we studied the model's behavior in two particular conditions. First, we investigated the relationship between model performance and the departure from isochrony in natural sentences. Our results show that natural sentences that are more temporally regular (less departure from isochrony) are still better recognized than those that are less regular. Second, we investigated the relationship between model performance and the naturalness considering isochronous sentences. Here again, our results show that the isochronous sentences temporally close to natural ones are better recognized than those with larger temporal distortion with respect to the natural timing, both in unit recognition (expected) and in event detection (unexpected). Overall, our simulation results with this implementation of the COSMO-Onset model partially replicate experimental results of Aubanel and Schwartz (2020). To the best of our knowledge, this was beyond the scope of previous computational models in the field.

2 COSMO-Onset model vs other computational models of speech perception

Altogether, how does our model COSMO-Onset compare to other computational models of speech perception in the literature? First, the clear difference between our model and psycholinguistics models such as TRACE and SHORTLIST resides in the separation between the *WHAT* and *WHEN* questions involved in speech perception (Arnal & Giraud, 2012), which is not a matter of concern to these classical psycholinguistic models. Second, with regard to other models coming from the neuroscience of speech perception research, the originality of our work resides in the design of a temporal control module with two mechanisms involved in syllabic segmentation, that is to say, a bottom-up onset detection mechanism

and a top-down temporal prediction mechanism. To add to the distinction between the COSMO-Onset model and the recent neuro-computational models of speech perception, it would be appropriate to refer to Marr's three-level taxonomy of cognitive models (Marr, 1982).

In this taxonomy (Marr, 1982), there are three types of cognitive models, hierarchically organized according to the type of constraint considered from the object of study under consideration. The first level is the **computational level**, which is only concerned with the description of the cognitive task to be solved. Models at this level are usually mainly concerned with describing how the cognitive task can be solved the best, and thus are concerned with optimality and rationality principles. In a sense, at this level, no specific cognitive system is considered, and ideal models developed at this level serve as asymptotic, "best-case scenario" reference. The second level is the **algorithm and representational level**, as the name suggests, which deals with the representations and algorithms hypothesized to solve the problem in a given cognitive system. The third and last level is the **implementation level**, concerned with hypothesized physical realization of the representations and algorithms in a given physical system that solves the cognitive task.

Clearly, with the COSMO-Onset model, we do not model nor simulate the physical realization of speech perception, which is rather the purpose of neuro-computational models. Instead, the COSMO-Onset model lies at the representational and algorithmic level of Marr's hierarchy. Its architecture is inspired by, and compatible with neuro-computational models based on neuroanatomy and neuronal measures on the one hand, and theories developed from experimental observations from the cognitive psychology of speech perception on the other hand. This abstraction from "implementation details" (how neurons and neuronal population precisely encode and exchange information) allows us to focus instead on the overall architecture of information encoding and manipulation in the system. The main hypothesis, in this framework, is that probabilities are the "common currency" in the system to represent previous knowledge and acquired sensory and perceptual information, which result in uncertain representations, and Bayesian inference is the tool to reason with such uncertain representations in the model. This approach provides modeling tools flexible enough to build complex model architectures, such as the one of the COSMO-Onset model while retaining the physical interpretability of model components. The methodology we applied to develop the COSMO-Onset model is embedded in the more general Bayesian Programming and Bayesian Algorithmic Modeling framework (Bessière et al., 2013; Diard, 2015).

3 Limitations and Perspectives

When designing the COSMO-Onset model, we aimed to elaborate and develop a fully Bayesian model of speech perception with a special focus on the temporal dimensions, that would be capable to simulate various experimental data. The first implementation of the model aimed mainly to illustrate the components of the model and study their interaction but only using synthetic stimuli as inputs. We identified the main limitations of this first version which are described in the discussion section of Chapter 3. In the previous chapter, we developed a second implementation to overcome some of those limitations. We moved from synthetic stimuli to real speech stimuli. However, considering the potential ambition of COSMO-Onset to be part of a general model of speech perception in humans, there are still, of course, many limitations to this latest and current version of the COSMO-Onset model, that could be extended in various ways and opened to new questions, as we will now discuss.

3.1 Addressing more extensively and realistically the role of higher levels in top-down temporal predictions

The temporal control module in COSMO-Onset is composed of two interacting parts: the bottom-up onset detection process and the top-down temporal prediction model. The focus in the present work was conceptual and exploratory, aiming at designing an architecture able to connect bottom-up processing and top-down predictions. The nature and content of bottom-up processing were rather carefully elaborated, thanks to the presentation of models and theories in Chapter 1. Still, of course, the content of top-down predictions per se is in the present state of our work quite preliminary.

In the first version in Chapter 3, top-down predictions are lexical. They are supposed to provide statistics about syllable duration within each word, which could be supposed to be part of the mental lexicon. In the second version in Chapter 6, the top-down model relies on prosodic information associating syllable duration with their position in a sentence. The structure of the mental lexicon on one hand, and the architecture of prosodic representations and processing on the other hand, are the object of an extremely rich literature that we do not claim to cover here, but which provides a number of avenues for later developments towards more realistic top-down predictive systems for our model. Of course, lexical and prosodic information could also be combined in various ways. Just to illustrate this, in the second version of COSMO-Onset presented in Chapter 6, we also tested informally different top-down models. For instance, with a term of the form $P(DSyl_{i+1} \mid type_i)$, the next syllable duration is predicted given the current syllable type decoded; with a term of the form $P(DSyl_{i+1} \mid type_i\ [pos = i])$,

syllable duration would be predicted using both the syllable type and the current syllable position.

Furthermore, both implementations of the COSMO-Onset model, and even the conceptual COSMO-Onset architecture, can be considered incomplete. For instance, they lack a number of processing levels, concerning for instance syntax and semantics/pragmatics. Incidentally, in both model variants, the model is not able to perform the recognition of a sequence of words. Of course, in ecological communication situations, words are not perceived independently without context, and context matters so that words embedded in sentences are more intelligible than the same words presented in isolation. A priority for future developments of COSMO-Onset would thus consist in extending its linguistic representations, by adding higher levels concerned with the processing and representation of syntax and semantics. Notice that, crucially, the effect of rhythm naturalness in sentence intelligibility in noise in the behavioral data by Aubanel and Schwartz (2020) could also be influenced by these lexical, prosodic, syntactic, and semantic levels, hence the importance to consider all of them in future developments.

We believe that the Bayesian modeling framework we use is actually favorable to consider and develop such extensions. Indeed, in previous versions of the COSMO family of models (Barnaud, Diard, et al., 2018), or in its close cousins related to speech production (Patri et al., 2016) we have been able to illustrate how model variants and model extensions could easily be integrated into unifying models, thanks to the versatility and interpretability of structured probabilistic models. In the domain of the study of visual word recognition and reading, such a strategy was also used to develop the BRAID family of models. The initial BRAID model was limited to simulating visuo-orthographic processes, and tasks such as letter perception, visual word recognition, and visual lexical decision (Ginestet et al., 2019; Phénix, 2018; Phénix et al., 2018). Extending the model with learning mechanisms yielded BRAID-Learn (Ginestet et al., 2022), extending it with phonological representations yielded BRAID-Phon (Saghiran et al., 2020), and these latter two were integrated, yielding BRAID-Acq, a model of reading acquisition, currently in development in Alexandra Steinhilber's Ph.D. thesis. Overall, the Bayesian framework for cognitive modeling (Bessière et al., 2013; Diard, 2015) that we applied to develop COSMO-Onset provides a way to explore and combine, step by step, progressive extensions of the COSMO architecture, as has been done over the years for the question of perceptuo-motor interactions in speech perception (Schwartz et al., 2022b).

3.2 Exploring the fusion models for real speech

With the current version of the COSMO-Onset model, we performed our simulation experiment using only the *AND* fusion operator, since we deemed it

more appropriate with respect to the data of the experiments, especially in the case of noisy conditions, and the apparent sensitivity of the ***RDF*** model. However, we have also proposed a second fusion operator, which is the *OR* model. Hence, it would be interesting to conduct the same studies but using the *OR* fusion operator to combine bottom-up onset detection and top-down temporal prediction. Considering the degraded conditions applied to the data, we expect the *AND* fusion model to perform better than the *OR* fusion model. Conversely, other experimental conditions could lead to favor the OR model (see Section 3.4).

At this stage, the question of the adequate fusion model efficiently combining bottom-up processing and top-town predictions for temporal segmentation remains open, and we already introduced some propositions about it in the discussion of Chapter 3, including exploring the possibility to select or combine various fusion models depending on the processing context (e.g. level of noise, speech style, etc).

At a more global level of questioning, it could also be questioned what happens if and when temporal segmentation per se becomes, in certain cases (e.g., in a very large amount of noise or strongly adverse conditions of communication), quite degraded and hence provide too many segmentation errors that would drive the decoding systems in great difficulty. In such cases where, in some sense, temporal segmentation could appear "counter-productive", it could be assumed that the decoding process operates alone without temporal control.

3.3 Embedding top-down temporal predictions into the neural oscillation framework

An important question for the cognitive neuroscience of speech processing concerns the way top-down information is neurally incorporated into the flow of information processing in the human brain. In this context, two points are worth mentioning. Firstly, the role of the beta band seems particularly relevant (Arnal, 2012; Arnal & Giraud, 2012; Hovsepyan et al., 2022; Pefkou et al., 2017; Poeppel & Assaneo, 2020; Sohoglu et al., 2012). An underlying question concerns the precise role of beta oscillations/synchronization in this process, to assess whether top-down information simply modulates the bottom-up activity in beta band frequencies (in that case, it would be a channel for conveying top-down information), or if it is oscillatory in nature. Since our model of top-down information and processing in COSMO-Onset does not incorporate any representation of neural oscillations, it is neutral with respect to this question, and the exact nature of beta synchrony processes remains an open issue.

Secondly, while we have abundantly discussed the role of the theta band in syllabic processing in Chapter 1, including higher layers of information processing and particularly at the prosodic and syntactic levels quite likely requires

considering the role of delta oscillations in the 1–2 Hz range, which is supposed to be the channel of information processing integrating higher linguistic units (Ding et al., 2017; Ghitza, 2011). While the chunking of words into sequences of syllables is done by the segmentation mechanisms relying on the theta band, the chunking of sentences into words would potentially be done through segmentation mechanisms using the delta band. Altogether, hence, the relationship between theta, delta, and beta oscillations in speech temporal processing in relation to bottom-up envelope modulations and top-down linguistic predictions shapes the agenda of future research in this field.

3.4 Testing other experimental paradigms

Our ambition at the beginning of this work was to use a number of experimental paradigms to assess our model. We finally focused on only one type of paradigm, concerned with noise, and the role of naturalness in sentences. However, a number of other paradigms for testing adverse conditions and the potential role of top-down temporal predictions exist, of course. Let us mention two of them.

Firstly, in line with the studies done and presented in Chapter 3, we would consider hypo-articulation conditions. Indeed, in such conditions, a number of real onsets would easily be missed from the signal alone, and it would be interesting to assess various conditions differing in the level of articulation (e.g., contrasting read speech, carefully articulated speech in e.g. a discourse or a talk, or highly under-articulated speech in conversations) to better assess the role of temporal predictions. According to the results presented in Chapter 3, we would expect the *OR* fusion operator to be a relevant model for hypo-articulation conditions.

Another experimental data that would be desirable to simulate concerns compressed speech. As shown in the work by Ghitza and Greenberg (2009), participants are able to cope with compressed speech up to a compression factor of 3. In other words, when the natural rhythm of a speaker is accelerated by a factor of 3, listeners are still able to more or less accurately understand the message. In their experiment using semantically unpredictable sentences composed of 6 to 8 words, they found that when inserting "appropriate" silence gaps, listeners could compensate for the compression hurdles, and still achieve good recognition performance. This illustrates the decoupling between the decoding module (which can process massively compressed signal) and the temporal control module (which is less robust to compression), which is a prominent feature of our model, inspired by these studies. We have not yet assessed whether and how the COSMO-Onset model would be able to account for such experimental data. This should be the topic of a future study.

More globally, a computational model such as COSMO-Onset can be claimed

to be relevant and/or useful in some sense, not only if it is shown to be able to address a single isolated problem or study such as the one we explored in Chapter 6, but also, more globally if it happens to resist to a number of consecutive experimental tests in relation with the literature. This point is discussed in very interesting terms in a recent paper by Blandón et al. (2021) in the context of speech development, and it sets the basis for future developments of the COSMO-Onset model.

3.5 The attention question

Throughout our work in this thesis, we have considered neural oscillations from the very specific perspective of neural entrainment. This stipulates that neuronal oscillations are primarily an intrinsic brain phenomenon that exists first by itself without any external excitation (Buzsáki & Draguhn, 2004) and that, in the case of external excitation, the rhythm generated by intrinsic oscillations adapts to closely track the rhythm present in the sensory stimulus. However, some studies go beyond this mechanistic description and propose to interpret neuronal oscillations in a theoretical framework related to attention processes (Calderone et al., 2014; Ward, 2003).

Usually, there are two frequency bands that are considered most likely to be involved in attention processes. Many studies have associated the gamma band oscillations with the modulation of bottom-up attention processes, whereas the beta band oscillations would be associated with top-down attention processes (Riddle et al., 2019). The exact mechanisms coordinating the way the different frequency bands communicate have yet to be elucidated and agreed upon. However, there is the general idea that the theta band oscillations, which are involved in processing sensory stimuli, also control gamma oscillations by ways of cross-frequency coupling (Canolty & Knight, 2010; Hyafil, Giraud, et al., 2015), where the phase of the theta band oscillations modulates the power of gamma oscillation.

Furthermore, there have been theoretical propositions relating cognitive disorders to abnormal neural entrainment. For example, it has been hypothesized that impaired neural entrainment in delta band oscillations (involved in prosody processing) and theta band oscillations (involved in syllable segmentation) would contribute to poor phonological processing in dyslexia (Goswami, 2011; Vidyasagar, 2019). In addition, some researchers have proposed a complementary view relating an auditory phonemic deficit to a gamma oscillation deficit (Giraud & Poeppel, 2012; Lehongre et al., 2013).

All the hypothesis we mentioned, as far as we understand them, are still subjects of debate. While further exploring them experimentally will certainly be fruitful, it could also be worth using computational models to address these

questions. The COSMO-Onset model may prove to be a useful tool in this regard, as it allows to carefully manipulate the interacting processes in the temporal control module.

This offers a tantalizing perspective. Indeed, if some cognitive disorders can be related to deficits in temporal processing characteristics, affecting either the bottom-up or top-down components of segmentation cues, and if these have been interpreted as related to attention processes, then this could mean that what we referred to as "temporal control" module in the COSMO-Onset model could also be interpreted as an "attentional" module. The common ground between both is the function of this module: indeed, both can be viewed as "filtering" and "controlling" the flow of information in the main decoding part of the model.

This further offers another perspective, that could possibly build a bridge to another, seemingly distant domain. Indeed, we wonder whether the way the temporal control module pilots acoustic processing in the COSMO-Onset framework (and more globally within all the neurocognitive models exploiting this concept, presented in Chapter 1) is related to so-called "attentional mechanisms" in recent deep-learning models. Attention in these models is related to the dynamic allocation of the focus of computation for a given task at a given time. To quote a seminal paper in the field (Bahdanau et al., 2014), "the decoder decides parts of the source sentence to pay attention to. By letting the decoder have an attention mechanism, we relieve the encoder from the burden of having to encode all information in the source sentence into a fixed-length vector". This corresponds actually quite closely to the role of the temporal control module in the present COSMO-Onset model. A perspective of the development of COSMO-Onset could be to attempt to better connect the neurophysiological concept of temporal control with the powerful computational concept of attention in deep neural networks – this is connected to the question that will be opened in the next and final section of this discussion.

4 A final word on deep learning models vs cognitive science models

Let us conclude this work by discussing a question that we find relevant. Considering the extraordinary development of machine learning tools and particularly deep learning models, and considering also the trend that the less constrained these models are, the better they can learn and the more efficient they happen to be, what is the sense to introduce architectural hypotheses on a specific temporal control device in a general speech processing architecture?

This actually leads to a much more general question, that is: within the realm of speech perception, can there be an opposition between deep learning

models and cognitive science models? Historically, deep learning research has been hugely impacted by cognitive science research in various topics, namely related to vision and language (text and speech). This is to contrast with the new tendency with the recent development of artificial intelligence, which seems to now open new directions of research in cognitive science, even though one can consider that they have been influencing each other for quite a long time now since both research fields have been around (Westberg et al., 2019). Recently, the fast pace of artificial intelligence system development has led to rather impressive outcomes, with some deep learning models achieving human-like performance or even outperforming it. Nevertheless, it is important to note that these models do not have the same goals as the cognitive science models, for the same tasks. This results in their use of holistic and sophisticated data processing techniques, ignoring and/or neglecting a whole lot of research from relevant cognitive science studies.

Now, with all the above said, why would one prefer one to the other? The "magical" answer is, of course, "it depends". For those concerned with high performance and industry applications, deep learning models ought to be the choice, whereas, for those of us concerned with understanding the functioning of the human brain, cognitive modeling is the way. And in between, we now have researchers who seek to combine the best of both worlds. A pivot in this domain is provided by Dupoux (2018), who proposed a roadmap for reverse-engineering the infant language-learner by using deep learning models with three main requirements. The first requirement is that models should be computationally scalable, by "[going] beyond conceptual and box-and-arrow frameworks", allowing them to deal with real data, which is the second requirement that Dupoux (2018) proposes. The author invites researchers to test their hypotheses using inputs as close as possible to infants' sensory signals. Finally, with the two previous requirements met, Dupoux (2018) proposes that models' performance ought to be compared to humans using what he called the "cognitive indistinguishability" that he defines as follows:

> *A human and a machine are cognitively indistinguishable with respect to a given set of cognitive tests when they yield numerically overlapping results when run on these tests.*

It is important to note here that the key element relies on the use of the word "test", making it very important to choose the appropriate experimental data on which to evaluate models.

Even though in our case, the COSMO-Onset model is not studying how infants learn language, it can be embedded within the proposed roadmap by Dupoux (2018), as is shown by the development of the model from the conceptual one to the latest variant dealing with real speech inputs. However, we want to

emphasize an important matter here. If seemingly all models, regardless of their architecture, can be considered successful in his proposal as long as the model performance compares with humans, we have another take, in which the model architecture does matter for us. Dupoux (2018) does not seem to distinguish between modeling and simulating, as long as the outputs are not far off. Following works by Maria (1997) and Rieder (2003), modeling can be defined as the process of producing models representing a system or an object, whereas, simulating can be defined as the operation of using a particular model to study the behavior of a system or an object. We took special care to design COSMO-Onset as first, a model of human speech perception, making explicit architectural choices informed by neuroscience and cognitive science studies on speech perception, that we then used to simulate specific tasks such as word recognition and syllabic event detection.

To be fair to deep learning models, it is becoming increasingly harder to argue against them when the sole consideration is model performance. They would undeniably outperform most, if not all of the cognitive science models, on many given tasks related to vision or speech perception. Nevertheless, the issue of the volume and cost of learning data and learning processes involved in these models is increasingly being highlighted, both for scientific reasons (massive learning à la GPT (T. Brown et al., 2020) or other deep learning models are cognitively unlikely and neurophysiologically impossible) and for sustainable development reasons (ecological considerations related to the consumption of computing energy and data storage). It thus seems that, for reasons of compromise between the performance and cost of the various models and parsimony, the cognitive modeling approach more generally, and the neuroscience-informed modeling in particular, is better suited than the deep learning approach. We believe that this is where models such as COSMO-Onset and other computational models may be useful for their cost-effectiveness and insights into human cognition.

Bibliography

Abercrombie, D. (1967). Elements of general phonetics, univ. *Press, Edinburgh* (cited on page 31).

Abercrombie, D. (1965). *Studies in phonetics and linguistics*. Oxford University Press. (Cited on page 31).

Afraimovich, V. S., Rabinovich, M. I., & Varona, P. (2004). Heteroclinic contours in neural ensembles and the winnerless competition principle. *International Journal of Bifurcation and Chaos*, *14*(04), 1195–1208 (cited on page 22).

Ahissar, E., & Ahissar, M. (2005). Processing of the temporal envelope of speech. *The auditory cortex* (pp. 313–332). Psychology Press. (Cited on page 10).

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences*, *98*(23), 13367–13372 (cited on page 37).

Ainsworth, M., Lee, S., Cunningham, M. O., Roopun, A. K., Traub, R. D., Kopell, N. J., & Whittington, M. A. (2011). Dual gamma rhythm generators control interlaminar synchrony in auditory cortex. *Journal of Neuroscience*, *31*(47), 17040–17051 (cited on pages 9, 24).

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation*, *9*(7), 1545–1588 (cited on page 84).

Amzica, F., & Steriade, M. (1998). Electrophysiological correlates of sleep delta waves. *Electroencephalography and clinical neurophysiology*, *107*(2), 69–83 (cited on page 9).

Arnal, L. H. (2012). Predicting "when" using the motor system's beta-band oscillations. *Frontiers in Human Neuroscience*, *6*, 225 (cited on pages 4, 10, 26, 29, 37, 125).

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390–398 (cited on pages 4, 11, 29, 37, 69, 121, 125).

Arnal, L. H., Poeppel, D., & Giraud, A.-l. (2015). Temporal coding in the auditory cortex. *Handbook of clinical neurology*, *129*, 85–98 (cited on page 10).

Arons, B. (1992). Techniques, perception, and applications of time-compressed speech. *Proceedings of 1992 Conference*, 169–177 (cited on page 15).

Attar, M., Mosleh, M., & Ansari-Asl, K. (2010). Isolated wordsrecognition based on random forest classifiers. *Proceedings of 2010 4th International Conference on Intelligent Information Technology* (cited on page 86).

Aubanel, V., Bayard, C., Strauß, A., & Schwartz, J.-L. (2020). The fharvard corpus: A phonemically-balanced French sentence resource for audiology and intelligibility research. *Speech Communication*, *124*, 68–74 (cited on pages 31, 74, 75, 86, 87, 92, 120).

Aubanel, V., Davis, C., & Kim, J. (2016). Exploring the role of brain oscillations in speech perception in noise: Intelligibility of isochronously retimed speech. *Frontiers in human neuroscience*, *10*, 430 (cited on page 32).

Aubanel, V., & Schwartz, J.-L. (2020). The role of isochrony in speech perception in noise. *Scientific Reports (Nature Publisher Group)*, *10*(1) (cited on pages i, iii, 3, 4, 31, 33–35, 37, 38, 70, 79, 82, 83, 92, 96, 100, 103, 104, 112–115, 119–121, 124).

Baars, B., & Gage, N. M. (2013). *Fundamentals of cognitive neuroscience: A beginner's guide*. Academic Press. (Cited on page 8).

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (cited on page 128).

Barnaud, M.-L., Bessière, P., Diard, J., & Schwartz, J.-L. (2018). Reanalyzing neurocognitive data on the role of the motor system in speech perception within COSMO, a Bayesian perceptuo-motor model of speech communication. *Brain & Language*, *187*, 19–32 (cited on page 5).

Barnaud, M.-L., Diard, J., Bessière, P., & Schwartz, J.-L. (2018). COSMO SylPhon: A Bayesian perceptuo-motor model to assess phonological learning. *Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech 2018)*, 3786–3790 (cited on page 124).

Başar, E., Başar-Eroglu, C., Karakaş, S., & Schürmann, M. (2001). Gamma, alpha, delta, and theta oscillations govern cognitive processes. *International journal of psychophysiology*, *39*(2-3), 241–248 (cited on page 9).

Başar, E., Başar-Eroğlu, C., Karakaş, S., & Schürmann, M. (1999). Are cognitive processes manifested in event-related gamma, alpha, theta and delta oscillations in the eeg? *Neuroscience letters*, *259*(3), 165–168 (cited on pages 8, 9).

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711 (cited on page 9).

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J.-M., Oostenveld, R., Dowdall, J. R., De Weerd, P., Kennedy, H., & Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*, *85*(2), 390–401 (cited on page 3).

Beckman, M. E. (1992). Evidence for speech rhythms across languages. *Speech perception, production and linguistic structure*, 457–463 (cited on page 10).

Benesty, J., Sondhi, M. M., Huang, Y., et al. (2008). *Springer handbook of speech processing* (Vol. 1). Springer. (Cited on page 1).

Bengio, Y. (1993). A connectionist approach to speech recognition. *Advances in pattern recognition systems using neural network technologies* (pp. 3–23). World Scientific. (Cited on page 83).

Berger, H. (1929). Über das elektroenkephalogramm des menschen [on the use of the encephalogram in humans]. *Archiv für psychiatrie und nervenkrankheiten*, *87*(1), 527–570 (cited on page 8).

Berwick, R. C., Beckers, G. J., Okanoya, K., & Bolhuis, J. J. (2012). A bird's eye view of human language evolution. *Frontiers in evolutionary neuroscience*, *4*, 5 (cited on page 21).

Berwick, R. C., Friederici, A. D., Chomsky, N., & Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. *Trends in cognitive sciences*, *17*(2), 89–98 (cited on page 1).

Bessière, P., Laugier, C., & Siegwart, R. (Eds.). (2008). *Probabilistic reasoning and decision making in sensory-motor systems* (Vol. 46). Springer. (Cited on pages 4, 159).

Bessière, P., Mazer, E., Ahuactzin, J. M., & Mekhnacha, K. (2013). *Bayesian programming*. CRC Press. (Cited on pages 42, 122, 124, 157, 169).

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227 (cited on page 84).

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer. (Cited on page 83).

Bispham, J. (2006). Rhythm in music: What is it? who has it? and why? *Music perception*, *24*(2), 125–134 (cited on page 8).

Blandón, M. A. C., Cristia, A., & Räsänen, O. (2021). Evaluation of computational models of infant language development against robust empirical data from meta-analyses: What, why, and how? (Cited on page 127).

Boë, L.-J., & Maeda, S. (1998). Modélisation de la croissance du conduit vocal. *Journées d'Études Linguistiques, La voyelle dans tous ses états*, 98–105 (cited on page 52).

Bolhuis, J. J., Okanoya, K., & Scharff, C. (2010). Twitter evolution: Converging mechanisms in birdsong and human speech. *Nature Reviews Neuroscience*, *11*(11), 747–759 (cited on page 21).

Borde, P., Kulkarni, S., Gawali, B., & Yannawar, P. (2020). Recognition of isolated digit using random forest for audio-visual speech recognition. *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences*, 1–8 (cited on page 86).

Börgers, C., Epstein, S., & Kopell, N. J. (2005). Background gamma rhythmicity and attention in cortical local circuits: A computational study. *Proceedings of the National Academy of Sciences*, *102*(19), 7002–7007 (cited on page 24).

Börgers, C., & Kopell, N. J. (2008). Gamma oscillations and stimulus selection. *Neural computation*, *20*(2), 383–414 (cited on page 24).

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140 (cited on page 84).

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32 (cited on pages 84, 85, 92).

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge. (Cited on page 85).

Brown, G. J., Cooke, M., & Mousset, E. (1996). Are neural oscillations the substrate of auditory grouping. *ESCA Tutorial and Workshop on the Auditory Basis of Speech Perception, Keele University, July*, 15–19 (cited on page 10).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901 (cited on page 130).

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: Experiences from the scikit-learn project. *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122 (cited on page 90).

Bunnell, H. T., Pennington, C., Yarrington, D., & Gray, J. (2005). Automatic personal synthetic voice construction. *Ninth European Conference on Speech Communication and Technology* (cited on page 16).

Burkitt, A. N. (2006). A review of the integrate-and-fire neuron model: I. homogeneous synaptic input. *Biological cybernetics*, *95*(1), 1–19 (cited on page 24).

Burnham, K., & Anderson, D. (2004). Model selection and multi-model inference. *A Practical Information-Theoretic Approach. Second. NY: Springer-Verlag*, *63*(2020), 10 (cited on page 69).

Buzsaki, G. (2006). *Rhythms of the brain*. Oxford University Press. (Cited on pages 2, 8, 9).

Buzsáki, G. (1998). Memory consolidation during sleep: A neurophysiological perspective. *Journal of sleep research*, *7*(S1), 17–23 (cited on page 9).

Buzsáki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science*, *304*(5679), 1926–1929 (cited on pages 2, 8, 9, 127).

Calderone, D. J., Lakatos, P., Butler, P. D., & Castellanos, F. X. (2014). Entrainment of neural oscillations as a modifiable substrate of attention. *Trends in cognitive sciences*, *18*(6), 300–309 (cited on pages 9, 11, 127).

Cannon, J., McCarthy, M. M., Lee, S., Lee, J., Börgers, C., Whittington, M. A., & Kopell, N. (2014). Neurosystems: Brain rhythms and cognitive processing. *European Journal of Neuroscience*, *39*(5), 705–719 (cited on pages 8, 9).

Canolty, R. T., Edwards, E., Dalal, S. S., Soltani, M., Nagarajan, S. S., Kirsch, H. E., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2006). High gamma power is phase-locked to theta oscillations in human neocortex. *science*, *313*(5793), 1626–1628 (cited on page 11).

Canolty, R. T., & Knight, R. T. (2010). The functional role of cross-frequency coupling. *Trends in cognitive sciences*, *14*(11), 506–515 (cited on pages 11, 127).

Cason, N., & Schön, D. (2012). Rhythmic priming enhances the phonological processing of speech. *Neuropsychologia*, *50*(11), 2652–2658 (cited on page 13).

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS computational biology*, *5*(7), e1000436 (cited on page 10).

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887–906 (cited on page 28).

Chinchor, N. (1992). Muc-4 evaluation metrics. *Proceedings of the Fourth Message Understanding Conference*, 22–29 (cited on page 20).

Clarke, J., Shelton, J., Venning, G., Hamer, J., & Taylor, S. (1976). The rhythm of the normal human heart. *The Lancet*, *308*(7984), 508–512 (cited on page 8).

Clayton, M. S., Yeung, N., & Kadosh, R. C. (2015). The roles of cortical oscillations in sustained attention. *Trends in cognitive sciences*, *19*(4), 188–195 (cited on page 9).

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. *Papers in laboratory phonology*, *1*, 283–333 (cited on page 26).

Cope, T. E., Sohoglu, E., Sedley, W., Patterson, K., Jones, P., Wiggins, J., Dawson, C., Grube, M., Carlyon, R., Griffiths, T., et al. (2017). Evidence

for causal top-down frontal contributions to predictive processes in speech perception. *Nature Communications*, *8*(1), 1–16 (cited on pages 4, 37).

Csibra, G., Davis, G., Spratling, M., & Johnson, M. (2000). Gamma oscillations and object processing in the infant brain. *Science*, *290*(5496), 1582–1585 (cited on page 9).

Cummins, F. (2012a). Looking for rhythm in speech. *Empirical Musicology Review*, *7* (cited on page 31).

Cummins, F. (2012b). Oscillators and syllables: A cautionary note. *Frontiers in psychology*, *3*, 364 (cited on page 31).

Cummins, F. (2015). Rhythm and speech. *The handbook of speech production*, 158–177 (cited on pages 10, 31).

Cutler, A. (1994). The perception of rhythm in language (cited on page 2).

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Norrmann, I. (2011). Explaining sonority projection effects. *Phonology*, *28*(2), 197–234 (cited on page 26).

Damineli, D. S., Portes, M. T., & Feijó, J. A. (2022). Electrifying rhythms in plant cells. *Current Opinion in Cell Biology*, *77*, 102113 (cited on page 8).

Dasher, R., & Bolinger, D. (1982). On pre-accentual lengthening. *Journal of the International Phonetic Association*, *12*(2), 58–71 (cited on page 31).

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of phonetics*, *11*(1), 51–62 (cited on page 31).

David, O., Kilner, J. M., & Friston, K. J. (2006). Mechanisms of evoked and induced responses in meg/eeg. *Neuroimage*, *31*(4), 1580–1591 (cited on page 8).

Davis, M. H., Gaskell, M. G., & Marslen-Wilson, W. (1998). Recognising embedded words in connected speech: Context and competition. *4th Neural Computation and Psychology Workshop, London, 9–11 April 1997*, 254–266 (cited on page 61).

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1-2), 132–147 (cited on pages 3, 36).

Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(1), 218 (cited on page 61).

Deco, G., & Corbetta, M. (2011). The dynamical balance of the brain at rest. *The Neuroscientist*, *17*(1), 107–123 (cited on page 16).

Diard, J. (2015). *Bayesian algorithmic modeling in cognitive science* (Habilitation à Diriger des Recherches (HDR)). Université Grenoble Alpes. (Cited on pages 122, 124, 157).

Dietterich, T. G. (2000a). Ensemble methods in machine learning. *International workshop on multiple classifier systems*, 1–15 (cited on page 84).

Dietterich, T. G. (2000b). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, *40*(2), 139–157 (cited on page 85).

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164 (cited on page 2).

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, *81*, 181–187 (cited on pages 2, 10, 98, 126).

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in human neuroscience*, *8*, 311 (cited on page 26).

Domingos, P. (2000). Bayesian averaging of classifiers and the overfitting problem. *ICML*, *747*, 223–230 (cited on page 84).

Donhauser, P. W., & Baillet, S. (2020). Two distinct neural timescales for predictive speech processing. *Neuron*, *105*(2), 385–393 (cited on page 3).

Dooling, R. (1992). Perception of speech sounds by birds. *Auditory physiology and perception* (pp. 407–413). Elsevier. (Cited on page 21).

Dooling, R. J., Leek, M. R., Gleich, O., & Dent, M. L. (2002). Auditory temporal resolution in birds: Discrimination of harmonic complexes. *The Journal of the Acoustical Society of America*, *112*(2), 748–759 (cited on page 21).

Dorman, M. F., Cutting, J. E., & Raphael, L. J. (1975). Perception of temporal order in vowel sequences with and without formant transitions. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(2), 121 (cited on page 52).

Doupe, A. J., & Kuhl, P. K. (1999). Birdsong and human speech: Common themes and mechanisms. *Annual review of neuroscience*, *22*(1), 567–631 (cited on page 21).

Drucker, H., Schapire, R., & Simard, P. (1992). Improving performance in neural networks using a boosting algorithm. *Advances in neural information processing systems*, *5* (cited on page 84).

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59 (cited on pages 129, 130).

Edin, F., Klingberg, T., Johansson, P., McNab, F., Tegnér, J., & Compte, A. (2009). Mechanism for top-down control of working memory capacity. *Proceedings of the National Academy of Sciences*, *106*(16), 6802–6807 (cited on page 36).

Engel, A. K., & Fries, P. (2010). Beta-band oscillations—signalling the status quo? *Current Opinion in Neurobiology*, *20*(2), 156–165 (cited on pages 4, 11).

Engel, A. K., & Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *Trends in cognitive sciences*, *5*(1), 16–25 (cited on page 2).

Ermentrout, G. B., & Kopell, N. (1986). Parabolic bursting in an excitable system coupled with a slow oscillation. *SIAM journal on applied mathematics*, *46*(2), 233–253 (cited on page 29).

Eyigöz, E., Gildea, D., & Oflazer, K. (2013). Multi-rate HMMs for word alignment. *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 494–502 (cited on page 4).

Fant, G. (1970). *Acoustic theory of speech production*. Walter de Gruyter. (Cited on page 52).

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, *15*(1), 3133–3181 (cited on page 85).

Ferreira, F. (1993). Creation of prosody during sentence production. *Psychological review*, *100*(2), 233 (cited on page 98).

Fiveash, A., Bedoin, N., Gordon, R. L., & Tillmann, B. (2021). Processing rhythm in speech and music: Shared mechanisms and implications for developmental speech and language disorders. *Neuropsychology*, *35*(8), 771 (cited on page 8).

Fiveash, A., Falk, S., & Tillmann, B. (2021). What you hear first, is what you get: Initial metrical cue presentation modulates syllable detection in sentence processing. *Attention, Perception, & Psychophysics*, *83*(4), 1861–1877 (cited on page 10).

Fontolan, L., Morillon, B., Liegeois-Chauvel, C., & Giraud, A.-L. (2014). The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. *Nature communications*, *5*(1), 1–10 (cited on page 3).

Foulke, E. (1971). The perception of time compressed speech. *Perception of Language* (cited on page 15).

Foulke, E., & Sticht, T. G. (1969). Review of research on the intelligibility and comprehension of accelerated speech. *Psychological bulletin*, *72*(1), 50 (cited on page 15).

Fries, P. (2015). Rhythms for cognition: Communication through coherence. *Neuron*, *88*(1), 220–235 (cited on pages 2, 9, 11).

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836 (cited on pages 3, 11, 35).

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221 (cited on pages 3, 11, 35).

Gales, M., & Young, S. (2008). The application of hidden markov models in speech recognition (cited on pages 2, 4).

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of experimental psychology: Human perception and performance*, *6*(1), 110 (cited on page 36).

Garofolo, J. S. (1993). Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993* (cited on pages 24, 30).

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: Bridging selective attention and working memory. *Trends in cognitive sciences*, *16*(2), 129–135 (cited on page 36).

Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, *2*, 130 (cited on pages 2, 7, 13, 18, 26, 35, 43, 45, 52, 126, 160).

Ghitza, O. (2013). The theta-syllable: A unit of speech information defined by cortical function. *Frontiers in Psychology*, *4*, 138 (cited on pages 16, 18).

Ghitza, O. (2020). "acoustic-driven oscillators as cortical pacemaker": A commentary on meyer, sun & martin (2019). *Language, Cognition and Neuroscience*, *35*(9), 1100–1105 (cited on pages 13, 14).

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, *66*(1-2), 113–126 (cited on pages 10, 15, 16, 126).

Gibbon, D., & Gut, U. (2001). Measuring speech rhythm. *Seventh European Conference on Speech Communication and Technology* (cited on page 10).

Gilbert, C. D., & Sigman, M. (2007). Brain states: Top-down influences in sensory processing. *Neuron*, *54*(5), 677–696 (cited on page 35).

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 (cited on pages 42, 159, 169).

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 (cited on pages 43, 124, 159).

Ginestet, E., Valdois, S., & Diard, J. (2022). Probabilistic modeling of orthographic learning based on visuo-attentional dynamics. *Psychonomic Bulletin & Review*, *29*, 1649–1672 (cited on page 124).

Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S., & Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron*, *56*(6), 1127–1134 (cited on pages 10, 16).

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511 (cited on pages 2, 7, 10, 16–18, 26, 35, 69, 127).

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 (cited on page 2).

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, *1*, 517–520 (cited on page 27).

Gómez, D. M., Berent, I., Benavides-Varela, S., Bion, R. A., Cattarossi, L., Nespor, M., & Mehler, J. (2014). Language universals at birth. *Proceedings of the National Academy of Sciences*, *111*(16), 5837–5841 (cited on page 26).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press. (Cited on page 83).

Goswami, U. (2011). A temporal sampling framework for developmental dyslexia. *Trends in cognitive sciences*, *15*(1), 3–10 (cited on page 127).

Goswami, U. (2019). Speech rhythm and language acquisition: An amplitude modulation phase hierarchy perspective. *Annals of the New York Academy of Sciences*, *1453*(1), 67–78 (cited on page 10).

Goswami, U. (2022). Language acquisition and speech rhythm patterns: An auditory neuroscience perspective. *Royal Society Open Science*, *9*(7), 211855 (cited on page 2).

Goswami, U., & Leong, V. (2013). Speech rhythm and temporal structure: Converging perspectives? *Laboratory Phonology*, *4*(1), 67–92 (cited on page 2).

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*, *7*(1982), 515–546 (cited on page 31).

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE international conference on acoustics, speech and signal processing*, 6645–6649 (cited on page 4).

Greenberg, S. (1998). A syllable-centric framework for the evolution of spoken language. *Behavioral and brain sciences*, *21*(4), 518–518 (cited on pages 2, 10).

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech—a syllable-centric perspective. *Journal of Phonetics*, *31*(3-4), 465–485 (cited on pages 2, 10, 98).

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint arXiv:2207.08815* (cited on page 85).

Grosjean, F., & Gee, J. P. (1987). Prosodic structure and spoken word recognition. *Cognition*, *25*(1-2), 135–155 (cited on page 2).

Haegens, S., Nácher, V., Hernández, A., Luna, R., Jensen, O., & Romo, R. (2011). Beta oscillations in the monkey sensorimotor network reflect somatosensory decision making. *Proceedings of the National Academy of Sciences*, *108*(26), 10708–10713 (cited on page 9).

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE intelligent systems*, *24*(2), 8–12 (cited on page 83).

Hamza, Y., Okalidou, A., Kyriafinis, G., & van Wieringen, A. (2018). Sonority's effect as a surface cue on lexical speech perception of children with cochlear implants. *Ear and Hearing*, *39*(5), 992–1007 (cited on page 26).

Hanslmayr, S., Gross, J., Klimesch, W., & Shapiro, K. L. (2011). The role of alpha oscillations in temporal attention. *Brain research reviews*, *67*(1-2), 331–343 (cited on page 9).

Harmony, T. (2013). The functional significance of delta oscillations in cognitive processing. *Frontiers in integrative neuroscience*, *7*, 83 (cited on page 9).

Harris, J. (2006). The phonology of being understood: Further arguments against sonority. *Lingua*, *116*(10), 1483–1494 (cited on page 26).

Heittola, T., Mesaros, A., Eronen, A., & Virtanen, T. (2013). Context-dependent sound event detection. *EURASIP Journal on Audio, Speech, and Music Processing*, *2013*(1), 1–13 (cited on page 20).

Hickok, G., Farahbod, H., & Saberi, K. (2015). The rhythm of perception: Entrainment to acoustic rhythms induces subsequent perceptual oscillation. *Psychological science*, *26*(7), 1006–1013 (cited on page 8).

Hidalgo, C., Pesnot-Lerousseau, J., Marquis, P., Roman, S., & Schön, D. (2019). Rhythmic training improves temporal anticipation and adaptation abilities in children with hearing loss during verbal interaction. *Journal of Speech, Language, and Hearing Research*, *62*(9), 3234–3247 (cited on page 13).

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views

of four research groups. *IEEE Signal processing magazine*, *29*(6), 82–97 (cited on page 2).

Hohmann, V. (2002). Frequency analysis and synthesis using a gammatone filterbank. *Acta Acustica united with Acustica*, *88*(3), 433–442 (cited on page 70).

Hohwy, J. (2017). Priors in perception: Top-down modulation, Bayesian perceptual learning rate, and prediction error minimization. *Consciousness and Cognition*, *47*, 75–85 (cited on page 3).

Holdsworth, J., Nimmo-Smith, I., Patterson, R., & Rice, P. (1988). Implementing a gammatone filter bank. *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, *1*, 1–5 (cited on page 26).

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, *79*(8), 2554–2558 (cited on page 22).

Hopfield, J. J., & Tank, D. W. (1985). "Neural" computation of decisions in optimization problems. *Biological cybernetics*, *52*(3), 141–152 (cited on page 22).

Hovsepyan, S., Olasagasti, I., & Giraud, A.-L. (2020). Combining predictive coding and neural oscillations enables online syllable recognition in natural speech. *Nature Communications*, *11*(1), 1–12 (cited on pages 2, 7, 18, 20, 28, 35, 36, 43, 45, 70, 89).

Hovsepyan, S., Olasagasti, I., & Giraud, A.-L. (2022). Rhythmic modulation of prediction errors: A possible role for the beta-range in speech processing. *bioRxiv* (cited on pages 36, 37, 125).

Huang, F., Xie, G., & Xiao, R. (2009). Research on ensemble learning. *2009 International Conference on Artificial Intelligence and Computational Intelligence*, *3*, 249–252 (cited on page 84).

Huggins, A. (1975). Temporally segmented speech. *Perception & Psychophysics*, *18*(2), 149–157 (cited on page 15).

Hyafil, A., & Cernak, M. (2015). *Neuromorphic based oscillatory device for incremental syllable boundary detection* (tech. rep.). Idiap. (Cited on page 28).

Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., & Giraud, A.-L. (2015). Speech encoding by coupled cortical theta and gamma oscillations. *eLife*, *4*, e06213 (cited on pages 2, 7, 18, 23–25, 28, 29, 35, 36, 43, 45, 51, 70).

Hyafil, A., Giraud, A.-L., Fontolan, L., & Gutkin, B. (2015). Neural cross-frequency coupling: Connecting architectures, mechanisms, and functions. *Trends in neurosciences*, *38*(11), 725–740 (cited on page 127).

Instruments, T. (1991). Ti 46-word speaker-dependent isolated word corpus. *Gaithersburg: NIST* (cited on page 22).

Jackendoff, R. (2003). Précis of foundations of language: Brain, meaning, grammar, evolution. *Behavioral and Brain Sciences*, *26*(6), 651–665 (cited on page 1).

Jadi, M. P., & Sejnowski, T. J. (2014). Cortical oscillations arise from contextual interactions that regulate sparse coding. *Proceedings of the National Academy of Sciences*, *111*(18), 6780–6785 (cited on page 24).

Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological review*, *83*(5), 323 (cited on pages 11, 12).

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological review*, *96*(3), 459 (cited on pages 11, 13).

Jones, M. R., Moynihan, H., MacKenzie, N., & Puente, J. (2002). Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological science*, *13*(4), 313–319 (cited on page 13).

Juang, B. H., & Chen, T. (1998). The past, present, and future of speech processing. *IEEE signal processing magazine*, *15*(3), 24–48 (cited on page 1).

Kamper, H., Jansen, A., & Goldwater, S. (2017). A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech & Language*, *46*, 154–174 (cited on page 3).

Kaya, E., & Henry, M. J. (2022). Reliable estimation of internal oscillator properties from a novel, fast-paced tapping paradigm. *Scientific reports*, *12*(1), 1–16 (cited on page 13).

Keil, J., & Senkowski, D. (2018). Neural oscillations orchestrate multisensory processing. *The Neuroscientist*, *24*(6), 609–626 (cited on page 8).

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS biology*, *16*(3), e2004473 (cited on page 11).

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863 (cited on page 3).

Kim, C., Gowda, D., Lee, D., Kim, J., Kumar, A., Kim, S., Garg, A., & Han, C. (2020). A review of on-device fully neural end-to-end automatic speech recognition algorithms. *arXiv preprint arXiv:2012.07974* (cited on page 4).

Kim, H.-C., & Ghahramani, Z. (2012). Bayesian classifier combination. *Artificial Intelligence and Statistics*, 619–627 (cited on page 84).

Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in cognitive sciences*, *16*(12), 606–617 (cited on page 9).

Klimesch, W., Doppelmayr, M., Russegger, H., Pachinger, T., & Schwaiger, J. (1998). Induced alpha band power changes in the human eeg and attention. *Neuroscience letters*, *244*(2), 73–76 (cited on page 8).

Knyazev, G. G., Savostyanov, A. N., & Levin, E. A. (2004). Alpha oscillations as a correlate of trait anxiety. *International journal of psychophysiology*, *53*(2), 147–160 (cited on page 9).

Kolinsky, R., Morais, J., & Cluytens, M. (1995). Intermediate representations in spoken word recognition; evidence from word illusions. *Journal of Memory and Language*, *34*(1), 19–40 (cited on page 2).

Kösem, A., Bosker, H. R., Takashima, A., Meyer, A., Jensen, O., & Hagoort, P. (2018). Neural entrainment determines the words we hear. *Current Biology*, *28*(18), 2867–2875 (cited on page 10).

Kösem, A., & Van Wassenhove, V. (2017). Distinct contributions of low-and high-frequency neural oscillations to speech comprehension. *Language, cognition and neuroscience*, *32*(5), 536–544 (cited on pages 3, 10).

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review*, *39*(4), 261–283 (cited on page 85).

Kotz, S. A., Ravignani, A., & Fitch, W. T. (2018). The evolution of rhythm processing. *Trends in cognitive sciences*, *22*(10), 896–910 (cited on page 8).

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *science*, *320*(5872), 110–113 (cited on page 9).

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological review*, *106*(1), 119 (cited on page 12).

Laufs, H. (2008). Endogenous brain oscillations and related networks detected by surface eeg-combined fMRI. *Human brain mapping*, *29*(7), 762–769 (cited on page 16).

Laurent, R., Barnaud, M.-L., Schwartz, J.-L., Bessière, P., & Diard, J. (2017). The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception. *Psychological Review*, *124*(5), 572–602 (cited on page 5).

Lebeltel, O., Bessière, P., Diard, J., & Mazer, E. (2004). Bayesian robot programming. *Autonomous Robots*, *16*(1), 49–79 (cited on page 157).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444 (cited on page 83).

Lee, T. S. (2002). Top-down influence in early visual processing: A Bayesian perspective. *Physiology & behavior*, *77*(4-5), 645–650 (cited on page 36).

Lehongre, K., Morillon, B., Giraud, A.-L., & Ramus, F. (2013). Impaired auditory sampling in dyslexia: Further evidence from combined fmri and eeg. *Frontiers in human neuroscience*, *7*, 454 (cited on page 127).

Lennes, M. (2009). Segmental features in spontaneous and read-aloud finnish. *Phonetics of Russian and Finnish* (cited on page 27).

Lindblom, B. (1968). Temporal organization of syllable production. *Quarterly progress and status report*, *9*(2-3), 1–5 (cited on page 98).

Lindblom, B., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *The Journal of the Acoustical Society of America*, *42*(4), 830–843 (cited on page 52).

Lippus, P., Tuisk, T., Salveste, N., & Teras, P. (2013). Phonetic corpus of Estonian spontaneous speech. *Institute of Estonian and General Linguistics, University of Tartu. DOI: https://doi. org/10.15155/TY. D* (cited on page 27).

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, *54*(6), 1001–1010 (cited on page 10).

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and brain sciences*, *21*(4), 499–511 (cited on page 10).

MacNeilage, P. F., & Davis, B. L. (2000). On the origin of internal structure of word forms. *Science*, *288*(5465), 527–531 (cited on page 10).

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. *Speech production and speech modelling* (pp. 131–149). Springer. (Cited on page 52).

Maïonchi-Pino, N., de Cara, B., Ecalle, J., & Magnan, A. (2012). Are French dyslexic children sensitive to consonant sonority in segmentation strategies? Preliminary evidence from a letter detection task. *Research in developmental disabilities*, *33*(1), 12–23 (cited on page 26).

Maki, J., & Beasley, D. (1976). Time and frequency altered speech. *Contemporary issues in experimental phonetics*, 419–457 (cited on page 15).

Marcus, S. M. (1981). Acoustic determinants of perceptual center (p-center) location. *Perception & psychophysics*, *30*(3), 247–256 (cited on pages 20, 75).

Maria, A. (1997). Introduction to modeling and simulation. *Proceedings of the 29th conference on Winter simulation*, 7–13 (cited on page 130).

Marr, D. (1982). A computational investigation into the human representation and processing of visual information. *WH San Francisco: Freeman and Company, San Francisco* (cited on page 122).

Martinet, A. (1960). Eléments de linguistique générale, colin, paris. New updated edition 1980 (cited on page 1).

Massaro, D. W., & Cohen, M. M. (1983). Phonological context in speech perception. *Perception & psychophysics*, *34*(4), 338–348 (cited on page 36).

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86 (cited on pages 1, 3, 5, 37, 43, 160).

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in cognitive sciences*, *10*(8), 363–369 (cited on page 3).

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological review*, *88*(5), 375 (cited on pages 1, 36, 43).

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). Librosa: Audio and music signal analysis in python. *Proceedings of the 14th python in science conference*, *8*, 18–25 (cited on page 89).

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *The Journal of the Acoustical Society of America*, *58*(4), 880–883 (cited on pages 24, 45).

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: State of the art and emerging mechanisms. *European Journal of Neuroscience*, *48*(7), 2609–2621 (cited on page 3).

Meynadier, Y. (2001). La syllabe phonétique et phonologique: Une introduction. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, *20*, 91–148 (cited on page 2).

Miller, J. L., & Eimas, P. D. (1995). Speech perception: From signal to word. *Annual Review of Psychology*, *46*, 467 (cited on page 1).

Mishra, S. K., & Lutman, M. E. (2014). Top-down influences of the medial olivocochlear efferent system in speech perception in noise. *PLoS One*, *9*(1), e85756 (cited on page 36).

Mitchell, T. M., & Mitchell, T. M. (1997). *Machine learning* (Vol. 1). McGraw-hill New York. (Cited on page 83).

Mitra, P. P., & Pesaran, B. (1999). Analysis of dynamic brain imaging data. *Biophysical journal*, *76*(2), 691–708 (cited on page 8).

Moratti, S., Clementz, B. A., Gao, Y., Ortiz, T., & Keil, A. (2007). Neural mechanisms of evoked oscillations: Stability and interaction with transient events. *Human brain mapping*, *28*(12), 1318–1333 (cited on page 8).

Morillon, B., Arnal, L. H., Schroeder, C. E., & Keitel, A. (2019). Prominence of delta oscillatory rhythms in the motor cortex and their relevance for

auditory and speech perception. *Neuroscience & Biobehavioral Reviews*, *107*, 136–142 (cited on page 9).

Morillon, B., Schroeder, C. E., Wyart, V., & Arnal, L. H. (2016). Temporal prediction in lieu of periodic stimulation. *Journal of Neuroscience*, *36*(8), 2342–2347 (cited on page 37).

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (p-centers). *Psychological review*, *83*(5), 405 (cited on pages 20, 32, 75).

Moulin-Frier, C., Diard, J., Schwartz, J.-L., & Bessière, P. (2015). COSMO ("Communicating about Objects using Sensory-Motor Operations"): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems. *Journal of Phonetics (special issue "On the cognitive nature of speech sound systems")*, *53*, 5–41 (cited on page 5).

Moulin-Frier, C., Laurent, R., Bessière, P., Schwartz, J.-L., & Diard, J. (2012). Adverse conditions improve distinguishability of auditory, motor and perceptuo-motor theories of speech perception: An exploratory Bayesian modeling study. *Language and Cognitive Processes*, *27*(7–8), 1240–1263 (cited on page 5).

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological cybernetics*, *66*(3), 241–251 (cited on pages 11, 35).

Murphy, K. (2002). *Dynamic Bayesian networks: Representation, inference and learning* (Ph. D. thesis). University of California, Berkeley. Berkeley, CA. (Cited on pages 4, 43).

Nabé, M., Diard, J., & Schwartz, J.-L. (2022). Isochronous is beautiful? Syllabic event detection in a neuro-inspired oscillatory model is facilitated by isochrony in speech. *Proc. Interspeech 2022*, 4671–4675 (cited on pages 25, 38, 73, 155).

Nabé, M., Schwartz, J.-L., & Diard, J. (2021). COSMO-Onset: A neurally-inspired computational model of spoken word recognition, combining top-down prediction and bottom-up detection of syllabic onsets. *Frontiers in Systems Neuroscience*, 75 (cited on pages 38, 39, 49, 82, 155, 157).

Nabé, M., Schwartz, J.-L., & Diard, J. (2022). Bayesian gates: A probabilistic modeling tool for temporal segmentation of sensory streams into sequences of perceptual accumulators. *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, 2257–2263 (cited on pages 38, 43, 155).

Nabé, Mamady. (2021). COSMO-Onset: A neurally-inspired computational model of word recognition. *Oral presentation in the Workshop on Bayesian models of cognition, language and speech organized by LPL and ILCB* (cited on page 156).

Nabé, Mamady. (2022). COSMO-Onset: A probabilistic word recognition model with top-down syllable duration knowledge involved in temporal segmentation of speech signals. *Oral presentation in the European Society for Cognitive Psychology conference (ESCOP2022)* (cited on page 156).

Nácher, V., Ledberg, A., Deco, G., & Romo, R. (2013). Coherent delta-band oscillations between cortical areas correlate with decision making. *Proceedings of the National Academy of Sciences*, *110*(37), 15085–15090 (cited on page 9).

Narayanan, N. S., & Laubach, M. (2006). Top-down control of motor cortex ensembles by dorsomedial prefrontal cortex. *Neuron*, *52*(5), 921–931 (cited on page 36).

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE access*, *7*, 19143–19165 (cited on page 2).

Nawas, K. K., Barik, M. K., & Khan, A. N. (2021). Speaker recognition using random forest. *ITM Web of Conferences*, *37*, 01022 (cited on page 86).

Nespor, M. (1990). On the rhythm parameter in phonology. *Logical issues in language acquisition*, *157*, 175 (cited on page 31).

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234 (cited on page 1).

Norris, D., & McQueen, J. M. (2008). Shortlist b: A Bayesian model of continuous speech recognition. *Psychological review*, *115*(2), 357 (cited on page 1).

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, *31*(1), 4–18 (cited on page 3).

Northoff, G., Qin, P., & Nakao, T. (2010). Rest-stimulus interaction in the brain: A review. *Trends in neurosciences*, *33*(6), 277–284 (cited on page 16).

Obin, N., Lamare, F., & Roebel, A. (2013). Syll-o-matic: An adaptive time-frequency representation for the automatic segmentation of speech into syllables. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6699–6703 (cited on page 20).

Obleser, J., Henry, M. J., & Lakatos, P. (2017). What do we talk about when we talk about rhythm? *PLoS biology*, *15*(9), e2002794 (cited on page 31).

Obleser, J., Herrmann, B., & Henry, M. J. (2012). Neural oscillations in speech: Don't be enslaved by the envelope. *Frontiers in human neuroscience*, *6*, 250 (cited on page 31).

Obleser, J., & Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. *Trends in cognitive sciences*, *23*(11), 913–926 (cited on page 13).

O'shaughnessy, D. (2000). *Speech communications: Human and machine* (Second edition). Wiley-IEEE Press. (Cited on page 1).

Paget, R. (2013). *Human speech: Some observations, experiments, and conclusions as to the nature*. Routledge. (Cited on page 1).

Parker, S. (2012). *The sonority controversy* (Vol. 18). Walter de Gruyter. (Cited on page 26).

Patel, A. D., & Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. *Cognition*, *87*(1), B35–B45 (cited on page 8).

Patel, A. D., Löfqvist, A., & Naito, W. (1999). The acoustics and kinematics of regularly timed speech: A database and method for the study of the p-center problem. *Proceedings of the 14th international congress of phonetic sciences*, *1*, 405–408 (cited on pages 20, 75).

Patri, J.-F., Diard, J., & Perrier, P. (2015). Optimal speech motor control and token-to-token variability: A Bayesian modeling approach. *Biological Cybernetics*, *109*(6), 611–626 (cited on page 5).

Patri, J.-F., Perrier, P., & Diard, J. (2016). Bayesian modeling in speech motor control: A principled structure for the integration of various constraints. *Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech 2016)*, 3588–3592 (cited on page 124).

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., & Rice, P. (1987). An efficient auditory filterbank based on the gammatone function. *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, *2*(7) (cited on page 26).

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., & Allerhand, M. (1992). Complex sounds and auditory images. *Auditory physiology and perception, proc. 9th international symposium on hearing* (pp. 429–446). Pergamon. (Cited on page 70).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830 (cited on page 90).

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in psychology*, *3*, 320 (cited on page 10).

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral cortex*, *23*(6), 1378–1387 (cited on page 37).

Pefkou, M., Arnal, L. H., Fontolan, L., & Giraud, A.-L. (2017). θ-band and β-band neural activity reflects independent syllable tracking and comprehension of time-compressed speech. *Journal of Neuroscience*, *37*(33), 7930–7938 (cited on pages 3, 36, 37, 125).

Pellegrino, F., Coupé, C., & Marsico, E. (2011). A cross-language perspective on speech information rate. *Language*, 539–558 (cited on pages 2, 10).

Perrone-Bertolotti, M., Kujala, J., Vidal, J. R., Hamame, C. M., Ossandon, T., Bertrand, O., Minotti, L., Kahane, P., Jerbi, K., & Lachaux, J.-P. (2012). How silent is silent reading? intracerebral evidence for top-down activation of temporal voice areas during reading. *Journal of Neuroscience*, *32*(49), 17554–17562 (cited on page 36).

Phénix, T. (2018). *Modélisation bayésienne algorithmique de la reconnaissance visuelle de mots et de l'attention visuelle* (Doctoral dissertation). Univ. Grenoble Alpes. (Cited on pages 42, 43, 124, 159, 167, 170).

Phénix, T., Valdois, S., & Diard, J. (2018). Reconciling opposite neighborhood frequency effects in lexical decision: Evidence from a novel probabilistic model of visual word recognition. In T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th annual conference of the cognitive science society* (pp. 2238–2243). Cognitive Science Society. (Cited on page 124).

Phillips, C. (2003). Linear order and constituency. *Linguistic inquiry*, *34*(1), 37–90 (cited on page 1).

Pisoni, D. B. (1985). Speech perception: Some new directions in research and theory. *The Journal of the Acoustical Society of America*, *78*(1), 381–388 (cited on page 1).

Pittman-Polletta, B. R., Wang, Y., Stanley, D. A., Schroeder, C. E., Whittington, M. A., & Kopell, N. J. (2021). Differential contributions of synaptic and intrinsic inhibitory currents to speech segmentation via flexible phase-locking in neural oscillators. *PLOS Computational Biology*, *17*(4), e1008783 (cited on page 70).

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, 1–13 (cited on pages 2, 10, 98, 125).

Price, P. J. (1980). Sonority and syllabicity: Acoustic correlates of perception. *Phonetica*, *37*(5-6), 327–343 (cited on page 26).

Przybyszewski, A. W. (1998). Vision: Does top-down processing help us to see? *Current Biology*, *8*(4), R135–R139 (cited on page 36).

Quené, H., & Port, R. F. (2005). Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica*, *62*(1), 1–13 (cited on page 13).

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286 (cited on pages 2, 70).

Rabiner, L. R., Lee, C.-H., Juang, B., & Wilpon, J. (1989). HMM clustering for connected word recognition. *International Conference on Acoustics, Speech, and Signal Processing,* 405–408 (cited on page 4).

Rabiner, L. R., Schafer, R. W. et al. (2007). Introduction to digital speech processing. *Foundations and Trends® in Signal Processing*, *1*(1–2), 1–194 (cited on page 1).

Rabinovich, M., Volkovskii, A., Lecanda, P., Huerta, R., Abarbanel, H., & Laurent, G. (2001). Dynamical encoding by networks of competing neuron groups: Winnerless competition. *Physical review letters*, *87*(6), 068102 (cited on page 22).

Radach, R., Huestegge, L., & Reilly, R. (2008). The role of global top-down factors in local eye-movement control in reading. *Psychological research*, *72*(6), 675–688 (cited on page 36).

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265–292 (cited on pages 2, 10, 31).

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87 (cited on pages 3, 11, 35).

Räsänen, O., Doyle, G., & Frank, M. C. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, *171*, 130–150 (cited on pages 2, 6, 7, 20, 25, 26, 35, 38, 45, 70, 73, 76–78, 81, 96, 120, 121).

Reiner, M., Rozengurt, R., & Barnea, A. (2014). Better than sleep: Theta neurofeedback training accelerates memory consolidation. *Biological psychology*, *95*, 45–53 (cited on page 9).

Riddle, J., Hwang, K., Cellier, D., Dhanani, S., & D'Esposito, M. (2019). Causal evidence for the role of neuronal oscillations in top–down and bottom–up attention. *Journal of cognitive neuroscience*, *31*(5), 768–779 (cited on page 127).

Rieder, W. G. (2003). Simulation and modeling. In R. A. Meyers (Ed.), *Encyclopedia of physical science and technology (third edition)* (Third Edition, pp. 815–835). Academic Press. (Cited on page 130).

Rimmele, J. M., Morillon, B., Poeppel, D., & Arnal, L. H. (2018). Proactive sensing of periodic and aperiodic auditory patterns. *Trends in Cognitive Sciences*, *22*(10), 870–882 (cited on pages 4, 11).

Roberts, J. W., Bennett, S. J., Elliott, D., & Hayes, S. J. (2014). Top-down and bottom-up processes during observation: Implications for motor learning. *European journal of sport science*, *14*(sup1), S250–S256 (cited on page 36).

Rohenkohl, G., & Nobre, A. C. (2011). Alpha oscillations related to anticipatory attention follow temporal expectations. *Journal of Neuroscience*, *31*(40), 14076–14084 (cited on page 9).

Rokach, L., & Maimon, O. (2005). Decision trees. *Data mining and knowledge discovery handbook* (pp. 165–192). Springer. (Cited on page 85).

Rosen, S. (1992). Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *336*(1278), 367–373 (cited on page 2).

Rothauser, E. (1969). Ieee recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics*, *17*, 225–246 (cited on page 31).

Saghiran, A., Valdois, S., & Diard, J. (2020). Simulating length and frequency effects across multiple tasks with the Bayesian model BRAID-Phon. *Proceedings of the 42th Annual Conference of the Cognitive Science Society*, 3158–3163 (cited on page 124).

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249 (cited on page 84).

Samuel, A. G. (2011). Speech perception. *Annual review of psychology*, *62*, 49–72 (cited on page 1).

Sasaki, Y. et al. (2007). The truth of the f-measure. 2007. *Manchester: School of Computer Science, University of Manchester* (cited on page 20).

Schapire, R. E. (1999). A brief introduction to boosting. *Ijcai*, *99*, 1401–1406 (cited on page 84).

Schön, D., & Tillmann, B. (2015). Short-and long-term rhythmic interventions: Perspectives for language rehabilitation. *Annals of the New York Academy of Sciences*, *1337*(1), 32–39 (cited on pages 13, 31).

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences*, *32*(1), 9–18 (cited on pages 9, 51).

Schwartz, J.-L., Barnaud, M.-L., Bessière, P., Georges, M.-A., Laurent, R., Moulin-Frier, C., Nabé, M., Patri, J.-F., Perrier, P., & Diard, J. (2022a). COSMO : Un modèle bayésien des fondements sensorimoteurs de la perception et de la production de la parole. *Actes des 34e Journées d'Études sur la Parole (JEP2022)*, 1033–1041 (cited on page 156).

Schwartz, J.-L., Barnaud, M.-L., Bessière, P., Georges, M.-A., Laurent, R., Moulin-Frier, C., Nabé, M., Patri, J.-F., Perrier, P., & Diard, J. (2022b). Cosmo: Un modèle bayésien des fondements sensorimoteurs de la perception et de la production de la parole. *34e Journées d'Études sur la Parole* (cited on page 124).

Schwartz, J.-L., Boë, L.-J., Badin, P., & Sawallis, T. R. (2012). Grounding stop place systems in the perceptuo-motor substance of speech: On the universality of the labial–coronal–velar stop series. *Journal of Phonetics*, *40*(1), 20–36 (cited on page 52).

Seliger, P., Tsimring, L. S., & Rabinovich, M. I. (2003). Dynamics-based sequential memory: Winnerless competition of patterns. *Physical Review E*, *67*(1), 011905 (cited on page 22).

Singer, W. (2001). Consciousness and the binding problem. *Annals of the New York Academy of Sciences*, *929*(1), 123–146 (cited on page 9).

Smith, A. (2006). Speech motor development: Integrating muscles, movements, and linguistic units. *Journal of communication disorders*, *39*(5), 331–349 (cited on page 1).

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *Journal of Neuroscience*, *32*(25), 8443–8453 (cited on pages 4, 11, 36, 125).

Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS computational biology*, *16*(10), e1008215 (cited on page 3).

Stevens, K. N., & Klatt, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *The Journal of the Acoustical Society of America*, *55*(3), 653–659 (cited on page 52).

Strauß, A., & Schwartz, J.-L. (2017). The syllable in the light of motor skills and neural oscillations. *Language, Cognition and Neuroscience*, *32*(5), 562–569 (cited on pages 32, 75).

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. *Proceedings of the IEEE international conference on computer vision*, 843–852 (cited on page 83).

Tallon-Baudry, C., & Bertrand, O. (1999). Oscillatory gamma activity in humans and its role in object representation. *Trends in cognitive sciences*, *3*(4), 151–162 (cited on page 8).

Temko, A., Nadeu, C., Macho, D., Malkin, R., Zieger, C., & Omologo, M. (2009). Acoustic event detection and classification. *Computers in the human interaction loop* (pp. 61–73). Springer. (Cited on page 20).

Terissi, L. D., Sad, G. D., Gómez, J. C., & Parodi, M. (2015). Audio-visual speech recognition scheme based on wavelets and random forests classification. *Iberoamerican Congress on Pattern Recognition*, 567–574 (cited on page 86).

Tillmann, B., Koelsch, S., Escoffier, N., Bigand, E., Lalitte, P., Friederici, A. D., & von Cramon, D. Y. (2006). Cognitive priming in sung and instrumental

music: Activation of inferior frontal cortex. *Neuroimage*, *31*(4), 1771–1782 (cited on page 13).

Tillmann, B., & Lebrun-Guillaud, G. (2006). Influence of tonal and temporal expectations on chord processing and on completion judgments of chord sequences. *Psychological Research*, *70*(5), 345–358 (cited on pages 13, 31).

Tzagarakis, C., Ince, N. F., Leuthold, A. C., & Pellizzer, G. (2010). Beta-band activity during motor planning reflects response uncertainty. *Journal of Neuroscience*, *30*(34), 11270–11277 (cited on page 9).

Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., & Lorenzi, C. (2017). A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America*, *142*(4), 1976–1989 (cited on page 98).

Vidyasagar, T. R. (2019). Visual attention and neural oscillations in reading and dyslexia: Are they possible targets for remediation? *Neuropsychologia*, *130*, 59–65 (cited on page 127).

Villing, R., Timoney, J., & Ward, T. E. (2004). Automatic blind syllable segmentation for continuous speech (cited on page 27).

Villing, R., Ward, T., & Timoney, J. (2006). Performance limits for envelope based automatic syllable segmentation. *2006 IET Irish Signals and Systems Conference*, 521–526 (cited on page 20).

Wang, D., & Narayanan, S. S. (2007). Robust speech rate estimation for spontaneous speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(8), 2190–2201 (cited on page 27).

Ward, L. M. (2003). Synchronous neural oscillations and cognitive processes. *Trends in cognitive sciences*, *7*(12), 553–559 (cited on pages 2, 11, 127).

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107 (cited on page 69).

Werker, J. F., & Tees, R. C. (1992). The organization and reorganization of human speech perception. *Annual Review of Neuroscience*, *15*(1), 377–402 (cited on page 1).

Westberg, M., Zelvelder, A., & Najjar, A. (2019). A historical perspective on cognitive science and its influence on xai research. *International workshop on explainable, transparent autonomous agents and multi-agent systems*, 205–219 (cited on page 129).

Wyart, V., De Gardelle, V., Scholl, J., & Summerfield, C. (2012). Rhythmic fluctuations in evidence accumulation during decision making in the human brain. *Neuron*, *76*(4), 847–858 (cited on pages 11, 51).

Xue, J., & Zhao, Y. (2008). Random forests of phonetic decision trees for acoustic modeling in conversational speech recognition. *IEEE transactions on audio, speech, and language processing*, *16*(3), 519–528 (cited on page 85).

Yildiz, I. B., & Kiebel, S. J. (2011). A hierarchical neuronal model for generation and online recognition of birdsongs. *PLoS Computational Biology*, *7*(12), e1002303 (cited on page 21).

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 Comput Biol*, *9*(9), e1003219 (cited on pages 2, 7, 21, 28, 29, 35, 36).

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (2002). The HTK book. *Cambridge university engineering department*, *3*(175), 12 (cited on pages 2, 19, 20, 105).

Yujian, L., & Bo, L. (2007). A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, *29*(6), 1091–1095 (cited on pages 20, 105).

Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., & Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. *Neuroimage*, *32*(4), 1826–1836 (cited on pages 3, 37).

Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. CRC press. (Cited on page 84).

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The Journal of the Acoustical Society of America*, *33*(2), 248–248 (cited on page 52).

Personal Bibliography

- Nabé, M., Schwartz, J.-L., & Diard, J. (2021). COSMO-Onset: A neurally-inspired computational model of spoken word recognition, combining top-down prediction and bottom-up detection of syllabic onsets. *Frontiers in Systems Neuroscience*, 75

- Nabé, M., Schwartz, J.-L., & Diard, J. (2022). Bayesian gates: A probabilistic modeling tool for temporal segmentation of sensory streams into sequences of perceptual accumulators. *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, 2257–2263

- Nabé, M., Diard, J., & Schwartz, J.-L. (2022). Isochronous is beautiful?

Syllabic event detection in a neuro-inspired oscillatory model is facilitated by isochrony in speech. *Proc. Interspeech 2022*, 4671–4675

- Schwartz, J.-L., Barnaud, M.-L., Bessière, P., Georges, M.-A., Laurent, R., Moulin-Frier, C., Nabé, M., Patri, J.-F., Perrier, P., & Diard, J. (2022a). COSMO : Un modèle bayésien des fondements sensorimoteurs de la perception et de la production de la parole. *Actes des 34e Journées d'Études sur la Parole (JEP2022)*, 1033–1041

Orals and Posters

- Nabé, Mamady. (2021). COSMO-Onset: A neurally-inspired computational model of word recognition. *Oral presentation in the Workshop on Bayesian models of cognition, language and speech organized by LPL and ILCB*
- Nabé, Mamady. (2022). COSMO-Onset: A probabilistic word recognition model with top-down syllable duration knowledge involved in temporal segmentation of speech signals. *Oral presentation in the European Society for Cognitive Psychology conference (ESCOP2022)*

Full COSMO-Onset first variant model specification

Note

This appendix is partially adapted from the supplementary materials of the published paper (Nabé et al., 2021).

To specify the full model (see Figure 3.1), we use the Bayesian Programming framework (Bessière et al., 2013; Diard, 2015; Lebeltel et al., 2004), which is a methodology for defining probabilistic models. In this methodology, a joint probability distribution is defined following three steps: first, all the relevant variables are listed and their domains are defined; second, the joint probability distribution is decomposed into a product of terms, and some of these are simplified thanks to conditional independence hypotheses; third and last, all terms of the decomposition need to be specified, and their parameters possibly identified from data using a learning mechanism. Once the joint probability distribution is fully defined in this manner, it can be used to "answer questions", that is to say, compute terms of interest by applying Bayesian inference. We now provide the complete definition of the model by following these four steps.

1 Variables

In our methodology, probabilistic variables are defined by their name and domains. The way we specify variable names, in our notation, deserves an introduction. Indeed, since the same representational space (e.g., the syllabic space) can be shared by several variables depending on their roles in the model (e.g., the perceived information or the lexically predicted information), then we compose variable names. For instance, SyP would be the perceived syllable and SyL would be the lexically predicted syllable. Furthermore, we use subscript indices to denote "position" in the speech sequence, and superscript indices to denote "time instants". For instance, SyP_2^{50} would be the probabilistic variable representing knowledge that the model has, at time instant 50, about the second syllable

perceived. Finally, we use a shorthand to denote sets of variables: $SyP_{1:N}^{1:T}$ is the set of all variables about perceived syllable variables, for all positions 1 to N, with $N = 3$ the number of syllable decoders and all time instants 1 to T, with T, arbitrarily set to 500, which is the longest word duration in the lexicon. The different variables of the model are as follows.

- $I_{1:12}^{1:T}$ represent the spectral contents of the acoustic signal Input (in the following, we capitalize the part of the variable definition which motivates its name). They take continuous values in the 2-dimensional space representing the first two formants F1, and F2 (in barks).

- $\Delta L_{1:12}^{1:T}$ represent the derivative of the Loudness of the acoustic signal. The loudness variable used to describe the stimulus, which is not represented inside the model, and thus has no probabilistic variables associated, takes continuous values; therefore, it is also the case for the $\Delta L_{1:12}^{1:T}$ variables.

- $Sil_{1:12}^{1:T}$ are binary variables derived from the loudness of the stimulus. They indicate "Silence" instants in the input by locating places where the loudness is null (i.e., 1 represents a silent time step, 0 otherwise).

- $FeP_{1:12}^{0:T}$ represent the set of possible phones (*Fe* for "features"), which is a discrete set of values:

$$Fe = \{a, i, u, p, t, @, \#\} \ ,$$

  where /@/ represents transition phones (acoustic features outside of the other categories) and /#/ is an end-of-sequence marker. $FeS_{1:12}^{1:T}$ and $FeL_{1:12}^{1:T}$ are defined over the same domain but used for different portions of the model: *FeP* variables represent Perceived features, *FeS* represent Sensed features and *FeL* represent Lexically predicted features.

- $SyP_{1:3}^{0:T}$ represent the set of possible Syllables, which is a discrete set of values:

$$Sy = /a/, /i/, /u/, /pa/, /pi/, /pu/, /ta/, /ti/, /tu/ \ .$$

  $SyS_{1:3}^{1:T}$ and $SyL_{1:3}^{1:T}$ are defined over the same domain and, as previously, for different portions of the model: *SyP* variables for Perceived syllables, *SyS* for Sensed syllables and *SyL* for Lexically predicted syllables.

- $WP^{0:T}$ represent the set of possible Words, which is a discrete set of values. All the words of the lexicon can be found in Table 3.1 of Chapter 3. $WS^{1:T}$ are defined over the same domain, and, as above, *WP* variables represent Perceived words and *WS* variables represent Sensed words.

- To connect different portions of the model, a set of so-called "coherence variables" (Bessière et al., 2008; Gilet et al., 2011) are defined; they are binary variables, taking values 0 or 1. They are all represented graphically identically as "λ" nodes in Figure 3.1, but they actually have different mathematical notations. For instance, $\lambda FeSP_{1:12}^{1:T}$ connect the sensed and perceived phone variables, $\lambda FePL_{1:12}^{1:T}$ connect the perceived and lexically predicted phones, and so on and so forth for $\lambda SySP_{1:3}^{1:T}$, $\lambda SyPL_{1:12}^{1:T}$ and $\lambda WSP^{1:T}$.

- $A_{1:15}^{1:T}$ are sets of so-called "control variables" (Ginestet et al., 2019; Phénix, 2018), which are Boolean variables. They are used to control the amount of information transferred through the coherence variables between the different representational layers. There are 15 sets of such control variables: 3 of them, $A_{13:15}^{1:T}$, control the quantity of information between the syllable lexical and perceptual layers and the other 12, $A_{1:12}^{1:T}$, control the quantity of information transferred between the feature sensory and perceptual layers; this mechanism to control information transfer is used to activate the different phone and syllable decoders sequentially.

- $OTD^{1:T}$, $OBU^{1:T}$, $OREF^{1:T}$ and $OC^{1:T}$ are Boolean variables, to represent the probability that there is a syllabic Onset event. The OTD variables represent the prediction of syllable onset events derived from word lexical knowledge, in a "Top-Down" manner; OBU represent syllable onset events detected from acoustic envelope processing, in a "Bottom-Up" manner; $OREF$ represent syllable onsets (more precisely, their absence thereof) during the REFractory period after the preceding onset; finally, OC represent the syllabic onsets resulting from the Combination of available information about these events (either from the OBU and $OREF$ variables in the "BU-only" variant of the model or from a fusion model with the OTD variables in the complete model).

2 Decomposition

We now consider the joint space described by the conjunction of all the variables we defined above. The joint probability distribution (JD) cannot, of course, be defined directly; instead, we decompose it into a product of terms and simplify them with conditional independence assumptions. This results in a dependency structure, which is graphically represented in Figure 3.1. In other words, the conditional independence assumptions correspond to the structural choices that result in the overall architecture of the model. These structural choices are broadly motivated by theoretical frameworks of the architecture, for instance

assuming a separation between the temporal control module and the decoding module. This assumption is shared with other models, such as the TEMPO model (Ghitza, 2011). Another example is the three-layer architecture separating phone, syllable and word levels, as a variant of the TRACE model (McClelland & Elman, 1986).

To implement these structural assumptions into the model, we have the following joint probability distribution (JD):

$$JD = P\left(\begin{array}{l} WP^{0:T}\ WS^{1:T} \\ SyL_{1:3}^{1:T}\, SyP_{1:3}^{0:T}\, SyS_{1:3}^{1:T} \\ FeL_{1:12}^{1:T}\ FeP_{1:12}^{0:T}\ FeS_{1:12}^{1:T} \\ \lambda WSP^{1:T}\ \lambda SyPL^{1:T}\ \lambda SySP^{1:T} \\ \lambda FeSP^{1:T}\ \lambda FePL^{1:T} \\ I_{1:12}^{1:T}\ \Delta L_{1:12}^{1:T}\ Sil_{1:12}^{1:T} \\ A_{1:15}^{1:T}\ OTD^{1:T}\ OBU^{1:T}\ OC^{1:T}\ OREF^{1:T} \end{array}\right)$$

that we decompose into:

$$JD = \left[\begin{array}{l} P(WP^0) \times \prod_{i=1}^{3} P(SyP_i^0) \times \prod_{j=1}^{12} P(FeP_j^0) \\ \times \prod_{t=1}^{T} \left[\begin{array}{l} P(A_{1:15}^t) \times P(OREF^t \mid A_{1:15}^t) \\ \times P(OC^t \mid OTD^t\, OBU^t\, OREF^t) \\ \times P(OTD^t \mid WS^t) \times P(WP^t \mid WP^{t-1}) \\ \times P(\lambda WSP^t \mid WS^t\ WP^t) \times P(WS^t) \\ \times \prod_{i=1}^{3} \left[\begin{array}{l} P(SyL_i^t \mid WS^t) \times P(\lambda SyPL_i^t \mid SyP_i^t\ SyL_i^t) \\ \times P(SyP_i^t \mid SyP_i^{t-1}) \times P(SyS_i^t) \\ \times P(\lambda SySP_i^t \mid SyS_i^t\ SyP_i^t\ A_i^t) \\ \times \prod_{j=4(i-1)}^{4i} \left[\begin{array}{l} P(FeL_j^t \mid SyS_i^t) \\ \times P(\lambda FePL_j^t \mid FeL_j^t\ FeP_j^t) \\ \times P(FeP_j^t \mid FeP_j^{t-1}) \\ \times P(\lambda FeSP_j^t \mid FeS_j^t\ FeP_j^t\ A_j^t) \\ \times P(I_j^t \mid FeS_j^t) \times P(FeS_j^t) \\ \times P(\Delta L_j^t) \times P(Sil_j^t \mid WS^t) \\ \times P(OBU^t \mid \Delta L_j^t) \end{array}\right] \end{array}\right] \end{array}\right] \end{array}\right]$$

Inside the temporal product ($\prod_{t=1}^{T} \ldots$), the first four terms along with the $P(OBU^t \mid \Delta L_j^t)$ in the innermost product, relate to the temporal control module of the COSMO-Onset model and the other terms relate to the decoding module. Terms related to the decoding module are organized "vertically" to match the structure of the graph representing the decoding module of the model, in Figure 3.1 (from the WP variables at the top to the stimulus variables at the bottom).

3 Parametric forms

We now define all the parametric forms of the probability distributions of the terms that appear in the decomposition of the joint probability distribution.

1. The prior probability distributions of the temporal perceptual models (that is to say, over WP, SyP and FeP), are all defined as discrete uniform probability distributions over their domains (resp., over words of the lexicon, syllables and phones): $\forall w, P([WP^0 = w]) = \frac{1}{|W|}$, $\forall s, i, P([SyP_i^0 = s]) = \frac{1}{|Sy|}$, $\forall f, j, P([FeP_j^0 = f]) = \frac{1}{|Fe|}$.

2. The dynamic probability distributions of the temporal perceptual models (that is to say, over WP, SyP and FeP), are all defined as discrete conditional probability distributions over their domains (resp., over words of the lexicon, syllables and phones). These are "quasi-Dirac" distributions, that is to say, they have almost probability 1 on their "diagonal", and a residual, non-zero probability everywhere else. For instance, for the phone perceptual dynamic model we note:

$$P([FeP_j^t = f^t] \mid [FeP_j^{t-1} = f^{t-1}]) = \begin{cases} \frac{1+leak_{Fe}}{1+|Fe|leak_{Fe}} & \text{if } f^t = f^{t-1}, \\ \frac{leak_{Fe}}{1+|Fe|leak_{Fe}} & \text{otherwise,} \end{cases}$$

   with $leak_{Fe} = 10^{-3}$. The dynamic models over syllables and words are defined in a similar manner, with parameters $leak_{Sy} = leak_W = 10^{-3}$. We note that this value is set empirically here; in the presented simulations, it mostly controls information decay speed, that is to say, how fast decoders return to their initial, uniform state in the absence of a stimulus. Such decays can be observed in portions of Figure 3.5 (e.g., the probability for syllable "pa" in the first decoder, in late iterations).

3. As in perceptual models, in sensory models, all probability distributions of the form $P(WS^t)$, $P(SyS_i^t)$ and $P(FeS_j^t)$ are defined by discrete uniform probability distributions over their respective domains.

4. $P(I_j^t \mid [FeS_j^t = f])$: for every phone f, the probability distribution over spectral contents of the acoustic signals is defined by a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ in the space of the first two formants F_1, F_2. Their parameters are given Table 1 for all the phones except the transitional phone /@/, which has a fixed probability value over the formant domain, arbitrarily set to 10^{-2}.

5. $P(FeL_j^t \mid [SyS_i^t = s])$: for every syllable s, the probability distribution over the phones it has at position j is a Dirac probability distribution, that is to

Table 1: Parameters of the Gaussian distributions over the spectral contents, in F_1, F_2 space, for the term $P(I_j^t \mid [FeS_j^t = f])$. For the vowels */a, i, u/*, the parameters are identified from the *VLAM* dataset (see Figure 2). Mean parameters are defined as vectors $\begin{pmatrix} F_1 & F_2 \end{pmatrix}$ and covariance matrices are defined as $\begin{pmatrix} var(F_1) & cov(F_1, F_2) \\ cov(F_1, F_2) & var(F_2) \end{pmatrix}$.

Phone	Mean (μ)	Covariance (Σ)
/a/	$\begin{pmatrix} 6.1148 & 8.8597 \end{pmatrix}$	$\begin{pmatrix} 0.001128 & -0.00199 \\ -0.00199 & 0.00962 \end{pmatrix}$
/i/	$\begin{pmatrix} 3.1784 & 11.2910 \end{pmatrix}$	$\begin{pmatrix} 0.01522 & -0.00578 \\ -0.00578 & 0.00248 \end{pmatrix}$
/u/	$\begin{pmatrix} 4.3340 & 7.0888 \end{pmatrix}$	$\begin{pmatrix} 0.0104 & 0.01124 \\ 0.01124 & 0.02427 \end{pmatrix}$
/p/	$\begin{pmatrix} 3.3362 & 8.0357 \end{pmatrix}$	$\begin{pmatrix} 0.0011 & -0.0019 \\ -0.0019 & 0.01217 \end{pmatrix}$
/t/	$\begin{pmatrix} 3.2598 & 9.6076 \end{pmatrix}$	$\begin{pmatrix} 0.00127 & -0.00204 \\ -0.00204 & 0.0097 \end{pmatrix}$
/#/	$\begin{pmatrix} 0.0 & 0.0 \end{pmatrix}$	$\begin{pmatrix} 0.001128 & -0.00199 \\ -0.00199 & 0.00962 \end{pmatrix}$

say, a discrete distribution with probability 1 for phone j, and probability 0 for other phones. We note:

$$P([FeL_j^t = f] \mid [SyS_i^t = s]) = \begin{cases} 1 & \text{if phone } f \text{ is at position } j \text{ of syllable } s, \\ 0 & \text{otherwise.} \end{cases}$$

6. $P(SyL_i^t \mid [WS^t = w])$: for every word w, the probability distribution over the syllables it has at position i is a Dirac probability distribution, that is to say, a discrete distribution with probability 1 for syllable i, and probability 0 for other syllables. We note:

$$P([SyL_i^t = s] \mid [WS^t = w]) = \begin{cases} 1 & \text{if syllable } s \text{ is at position } i \text{ of word } w, \\ 0 & \text{otherwise.} \end{cases}$$

7. Perceptual and lexical variables are connected by coherence variables, so that the terms associated, by definition, are specified by:

$$P([\lambda SyPL_i^t = 1] \mid [SyP_i^t = s_p]\ [SyL_i^t = s_l]) = \begin{cases} 1 & \text{if } s_l = s_p \\ 0 & \text{otherwise.} \end{cases}$$

$$P([\lambda FePL_j^t = 1] \mid [FeP_j^t = f_p]\ [FeL_j^t = f_l]) = \begin{cases} 1 & \text{if } f_l = f_p \\ 0 & \text{otherwise.} \end{cases}$$

$P([\lambda WSP^t = 1] \mid [WS^t = w_s]\ [WP^t = w_p])$ is defined in the same manner:

$$P([\lambda WSP^t = 1] \mid [WS^t = w_s]\ [WP^t = w_p]) \quad = \quad \begin{cases} 1 & \text{if } w_s = w_p \\ 0 & \text{otherwise} \end{cases}$$

Technical details about the properties deriving from this definition of coherence variables are found in Appendix 5.

8. Sensory and perceptual variables are connected by controlled coherence variables, so that the terms associated, by definition, are specified by:

$$P([\lambda SySP_i^t = 1] \mid [SyS_i^t = s_s]\ [SyP_i^t = s_p]\ [A^t = a_s])$$
$$= \begin{cases} 1 & \text{if } s_s = s_p \text{ and } a_s = 1 \\ 0 & \text{if } s_s \neq s_p \text{ and } a_s = 1 \\ 1/|Sy| & \text{if } a_s = 0 \end{cases}$$
$$P([\lambda FeSP_j^t = 1] \mid [FeS_j^t = f_s]\ [FeP_j^t = f_p]\ [A^t = a_s])$$
$$= \begin{cases} 1 & \text{if } f_s = f_p \text{ and } a_s = 1 \\ 0 & \text{if } f_s \neq f_p \text{ and } a_s = 1 \\ 1/|Fe| & \text{if } a_s = 0. \end{cases}$$

For both expressions, the value of a_s controls the sequential activation of the corresponding decoders. At any given time step, only one syllabic decoder is activated, with $a_s = 1$ for this decoder, and $a_s = 0$ for others. The same applies to phonetic decoders within syllabic decoders. Technical details about the properties deriving from this definition of controlled coherence variables are found in Appendix 5.

9. $P(A_{1:15}^t)$ are defined as discrete prior probability distributions, chosen as a function of a decision process applied to the result of inference over variable OC^{t-1} (see below, Appendix 4). If this computation finds that the probability that OC^{t-1} is *True* is larger than a decision threshold τ (set empirically to 0.4 in the reported experiments), then, for sequencing phone decoders, the current decoder i is closed (by setting the probability $P([A_i^t = 1])$ to 0), and the next decoder $i + 1$ is opened (by setting the probability $P([A_{i+1}^t = 1])$ to a small, non-zero value α_{Fe}). A similar mechanism is employed for sequencing syllable decoders.

10. $P(OBU^t \mid \Delta L_j^t)$ is defined with a scaled logistic function:

$$P([OBU^t = \mathit{True}] \mid [\Delta L_j^t = \delta_L]) = \begin{cases} 2 \times \mathit{logistic}(\mathit{uphillC}) - 1 & \text{if} \quad \delta_L \geq 0 \\ 0 & \text{else} \end{cases}$$

where *uphillC* is a discrete counter, used to count the number of time steps

where energy in the signal envelope was increasing, *logistic* is the logistic function,

$$logistic(x) = \frac{1}{1 + \exp\left(\frac{-(x-\mu)}{s}\right)} ,$$

with μ and s parameters, respectively the mean and a scale parameter proportional to the standard deviation. For all the simulations presented, their values are respectively set to 0 and 1.

11. $P(\Delta L_j^t)$ are probability distributions of the derivatives of the loudness of the stimulus; they are defined as uniform probability distributions (in practice, since values for variables ΔL_d^t are provided by the stimulus, this choice is arbitrary and without consequence).

12. The term $P(Sil_j^t \mid WS^t)$ is defined by:

$$P(Sil_j^t = 1 \mid WS^t = w_s) = logistic(duration(w_s)) ,$$

with *logistic* the same logistic function, with the same parameters, as in the $P(OBU^t \mid \Delta L_j^t)$ term, and $duration(w_s)$ a function that provides, for each known word, its duration in time steps according to the lexicon (see Table 3.1).

13. The term $P(OTD^t \mid WS^t)$ is defined by:

$$P([OTD^t = True] \mid [WS^t = w_s]) = \sum_{s \in w_s} \mathcal{N}(\mu = TimeOnset(s, w_s), \sigma^2)$$

where $\mathcal{N}$ is the Normal probability density function, and $TimeOnset(s, w_s)$ is a function providing the time-instant at which the syllabic onset of syllable s in word w_s is expected. For all the presented simulations, the variance σ^2, controlling the dispersion of the temporal windows in which onsets are expected, is set to an arbitrary value of 10. One "normal kernel" is associated to predict each syllabic onset in a word, and they are summed to form the probability of onsets over all time steps. Figure 1 shows an illustration of the resulting probability profile, over all time steps, for the word *"pata"*. Since it is a bi-syllabic word, there are two normal kernels, around time steps 0 and 150.

14. The $P(OREF^t \mid A_{1:15}^t)$ term implements the refractory period, that is to say, it sets the probability to have another onset event to 0 for the next 50 time steps after the last detected onset. Technically, it is a conditional Dirac probability distribution, so that $P([OREF^t = True] \mid A_{1:15}^t)$ is 0 if the last time-instant at which $A_{1:15}^t$ was *True* is not yet "old enough" (this

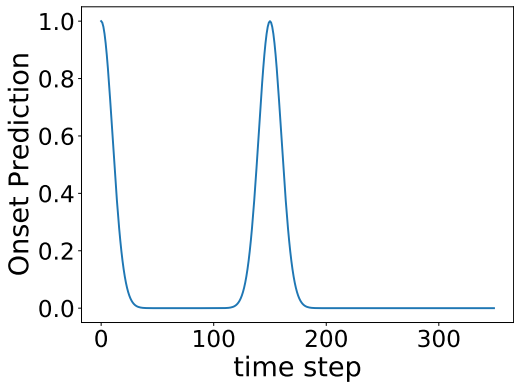

Figure 1: Top-down temporal prediction of syllabic onsets for the word *"pata"*, in the term $P([OTD^t = True] \mid [WS^t = pata])$. On the x-axis, the simulated time steps, and on the y-axis, the probability of predicted onsets at different time steps.

relies on an internal counter to describe this "memory", which is technically "outside" of the probabilistic description of the model, for simplicity).

15. The final term, $P(OC^t \mid OTD^t\ OBU^t\ OREF^t)$, defines the fusion operator. In the *AND* variant of the fusion operator, we define

$$\begin{aligned} &P([OC^t = T] \mid [OTD^t = T]\ [OBU^t = T]\ [OREF^t = T]) \\ &\quad = \quad P([OTD^t = T]) \times P([OBU^t = T]) \times P([OREF^t = T])\ , \end{aligned}$$

with T for the *True* Boolean value. This implements a combination in which the probability that there is onset is the product of the probability of the three "temporal submodels" (the top-down prediction, sensory detection, and refractory period). As a consequence of this product, the probability value can be close to one only when the three components agree and also provide a high probability that there is an onset; this explains why we denote this an "*AND*" combination. To define the *OR* operator, we apply De Morgan's law, $A \vee B = \overline{\overline{A} \wedge \overline{B}}$, and define:

$$\begin{aligned} &P([OC^t = T] \mid [OTD^t = T]\ [OBU^t = T]\ [OREF^t = T]) \\ &\quad = \quad \big(1 - (1 - P([OTD^t = T])) \times (1 - P([OBU^t = T]))\big) \times P([OREF^t = T])\ . \end{aligned}$$

Therefore, notice that, since the *AND* or *OR* fusion between bottom-up and top-down event detection is applied before the refractory process occurs, the *AND* operator could be construed as a probabilistic implementation of "*OTD AND OBU AND OREF*", while the *OR* operator could be noted as a probabilistic "(*OTD OR OBU*) *AND OREF*".

4 Inference for simulating word recognition

Here, we detail how word recognition and onset detection are computed in the model. Both correspond to probabilistic computations, computed in an online manner, thanks to recursive solutions provided by Bayesian inference. Both word recognition and onset detection are thus computed at each time step: word recognition proceeds assuming the states of the phone and syllable decoders as given, and onset detection, informed by word recognition, proceeds to compute the states of phones and syllables decoders for the next time step. In other words, model simulation proceeds in an iterative manner, as only probability distributions at time t are needed to compute probability distributions at time $t+1$ (even though for visualization purposes, we also memorize the whole history of probability distributions, this is not required for simulations).

Consider first the word decoding. Formally, word decoding relies on phone decoding and syllable decoding. To simulate these, we compute the probability distributions over the perceived phones FeP, syllables SyP and word WP, at each time step, assuming that the stimulus and states of each phone and syllable decoders (i.e., whether they are active or not) are given. To differentiate these three computations, we use the coherence variables to limit the propagation of information extracted from the stimulus into the model.

Consider first phone decoding. We are thus interested in computing, $\forall j, t$, $\mathcal{Q}Fe_j^t = P(FeP_j^t \mid I_j^{0:t}[\lambda FeSP_j^{0:t} = 1])$. Applying Bayesian inference in the model yields:

$$\mathcal{Q}Fe_j^t \propto \left(\begin{array}{l} [\alpha_{Fe} P(I_j^t \mid FeS_j^t) + (1-\alpha_{Fe})\mathcal{U}_{Fe}] \\ \times \sum_{FeP_j^{t-1}} P(FeP_j^t \mid FeP_j^{t-1})\ \mathcal{Q}Fe_j^{t-1} \end{array} \right), \tag{1}$$

where α_{Fe} is either equal to a constant (set to 10^{-1} for the simulations) when the corresponding phonetic decoder is activated, or 0 otherwise, and $\mathcal{U}_{Fe}$ is the uniform probability value over the phone space.

In a similar manner, for syllable decoding, we compute $\mathcal{Q}Sy_i^t = P(SyP_i^t \mid I_J^{0:t}[\lambda SySP_i^{0:t} = 1]\ [\lambda FeSP_J^{0:t} = 1])$ (with J denoting the set of subscripts from $4(i-1)+1$ to $4i$). Applying Bayesian inference yields:

$$\mathcal{Q}Sy_i^t \propto \left(\begin{array}{l} \prod_{j=4(i-1)+1}^{4i} [\alpha_{Sy} \langle P(FeL_j^t \mid SyS_i^t), \mathcal{Q}Fe_j^t \rangle + (1-\alpha_{Sy})\mathcal{U}_{Sy}] \\ \times \sum_{SyP_i^{t-1}} P(SyP_i^t \mid SyP_i^{t-1})\ \mathcal{Q}Sy_i^{t-1} \end{array} \right), \tag{2}$$

where α_{Sy} is either equal to a constant (set to 10^{-2} for the simulations) when the corresponding syllabic decoder is activated, or 0 otherwise, and $\mathcal{U}_{Sy}$ is the uniform probability value over the syllable space.

Finally, for word decoding, we compute $\mathcal{Q}W^t = P(WP^t \mid I_{1:12}^{0:t}\ Sil_{1:12}^{0:t}\ [\lambda WSP^{0:t} =$

1] $[\lambda SySP^{0:t}_{1:3} = 1]$ $[\lambda FeSP^{0:t}_{1:12} = 1]))$. Applying Bayesian inference yields:

$$\mathcal{Q}W^t \propto \left(\begin{array}{l} P(Sil \mid WS^t) \\ \times \prod_{i=1}^{3} [\alpha_W \langle P(SyL_i^t \mid WS^t), \mathcal{Q}Sy_i^t \rangle + (1 - \alpha_W)\mathcal{U}_W] \\ \times \sum_{WP^{t-1}} P(WP^t \mid WP^{t-1})\ \mathcal{Q}W^{t-1} \end{array} \right) , \qquad (3)$$

where α_W is equal to a constant (set to 10^{-2} for the simulations) and $\mathcal{U}_W$ is the uniform probability value over the word space.

We note that these inferences are approximate inferences. First, we do not take into account the feedback loops required by the complete, loopy dependency structure of the probabilistic model (due to the different Markov chains in parallel). Indeed, even though it is represented as a tree in Figure 3.1, the dependency structure of the decoding module contains variables with self-loops: in other words, the decoding module is a set of Markov chains, one over each such variable, interacting at each time step through the dependency structure shown in Figure 1. In such a hierarchical dynamic model, exact Bayesian inference would require sophisticated techniques, and even approximate Bayesian inference would require feed-forward and feedback information propagation until numerical convergence. Here, we proceed in a single-pass forward inference, as a first, rough approximation. (However, we note that introducing a feedback pass in a model with a similar architecture (Phénix, 2018) enabled contextual effects to appear in decoding; in our case, this would provide context effects of word recognition for syllable-decoding, and of syllable-recognition for phone-decoding. These effects are outside of the scope of the current model.) Second, we also do not consider the information propagation to the temporal module, and consider inference in the decoding portion of the model as independent (it "receives", from the temporal module, only onset events, and not probability distributions over onsets; in other words, the temporal module is seen, by the decoding module, as an external, independent sensor providing the states (open or closed) of phone and syllable decoders).

Consider, second, onset detection. At each time step, once word decoding is computed, we then compute, in the temporal module, onset detection, to update the states of phone and syllable decoders for the next, upcoming time step. Inference for the bottom-up, sensory detection of onsets simply proceeds by referring to the $P(OBU^t \mid \Delta L_j^t)$ term. On the other hand, for the top-down, prediction of onsets, we compute the probability:

$$P([OTD^t = True]) = \sum_{ws^t} P([OTD^t = True] \mid [WS^t = ws^t])\ \mathcal{Q}W^t \ .$$

In other words, we compute the probability that there would be an onset, according to the lexical models of all words, simultaneously, but weighed according to

the current probability distribution over words as computed by word recognition.

Computations of the probabilities of onsets in the refractory model, and in the fusion model, simply proceed by applying the corresponding definitions of their probabilistic terms. Considering, for instance, the *AND* fusion operator, we thus compute:

$$\begin{aligned} & P([OC^t = True] \mid \Delta L^t_{1:12}\ Sil^t_{1:12}\ I^t_{1:12}[\Lambda = 1]) \\ = & \left(\sum_{ws^t} P([OTD^t = True] \mid [WS^t = ws^t])\ \mathcal{QW}^t \right) \\ & \times P([OBU^t = T] \mid \Delta L^t_j) \times P([OREF^t = T])\ . \end{aligned}$$

with j the index of the currently active phone decoder, and Λ representing the set of all coherence variables of the decoding module.

The final step of onset detection is to apply the decision process on the computed probability distribution: when the probability that $[OC^t = True]$ is above a threshold, an onset is considered to be detected, which updates the states of phone and syllable decoders. Technically, this is done by changing the prior distributions over $P(A^{t+1}_{1:15})$ (see Appendix 3). This final step is not properly a "probabilistic dependency" in the model; that is why it is represented as a dotted arrow in Figure 1.

5 Using coherence and controlled coherence variables for controlling decoder input

Here, we detail how coherence variables and controlled coherence variables, can be used to control, in the model, when decoders are fed with sensory input. To do so, we consider a small portion of the model around the first phone decoder (without loss of generality, as this would apply to other phone decoders and to syllable decoders, as well).

Consider thus the first phone decoder. To recall, it is essentially a temporal model over variables $FeP^{0:T}_1$, defined by a dynamic model $P(FeP^t_1 \mid FeP^{t-1}_1)$. To simplify, we consider it provides, at time t, a distribution over perceived phones, noted here $P(FeP^t_1)$. This temporal model is fed sensory information from the input; this involves an inversion of the term $P(I^t_j \mid FeS^t_1)$. To simplify, we consider it provides, at time t, a distribution over sensed phones, noted here $P(FeS^t_1)$. We connect these distributions with a simple coherence variable $\lambda FeSP^t_1$. Therefore, we consider the model $P(FeP^t_1\ FeS^t_1\ \lambda FeSP^t_1)$, defined by:

$$P(FeP^t_1\ FeS^t_1\ \lambda FeSP^t_1) = P(FeP^t_1)P(FeS^t_1)P(\lambda FeSP^t_1 \mid FeP^t_1\ FeS^t_1)\ ,$$

with the term $P(\lambda FeSP_1^t \mid FeP_1^t\ FeS_1^t)$ defined by:

$$P([\lambda FeSP_1^t = 1] \mid [FeP_1^t = f_p]\ [FeS_1^t = f_s]) \quad = \quad \begin{cases} 1 & \text{if } f_p = f_S \\ 0 & \text{otherwise.} \end{cases}$$

We now demonstrate that, in this simplified portion of the model, coherence variable $\lambda FeSP_1^t$ can be employed as a "Bayesian switch", that is, we can choose during inference, whether information propagates from sensory information about phones to the phone decoder, or not. These demonstrations are adapted from other texts about coherence variables as Bayesian switches (Bessière et al., 2013; Gilet et al., 2011).

First, consider computing $P(FeP_1^t)$ in the model as defined above. The result of Bayesian inference can be shown to be equal to $P(FeP_1^t)$, since it appears as is in the decomposition of the joint probability distribution. By assumption, $P(FeP_1^t)$ is thus independent of the sensory distribution $P(FeS_1^t)$. Here, the coherence variable is unspecified, and this can be interpreted as "opening" the Bayesian switch. In other words, whatever information is in $P(FeS_1^t)$, it does not affect $P(FeP_1^t)$.

Second, consider computing $P([FeP_1^t = f_p] \mid [\lambda FeSP_1^t = 1])$:

$$\begin{aligned} & P([FeP_1^t = f_p] \mid [\lambda FeSP_1^t = 1]) \\ & \quad \propto \sum_{FeS_1^t} P([\lambda FeSP_1^t = 1]\ [FeP_1^t = f_p]\ FeS_1^t) \\ & \quad \propto \sum_{FeS_1^t} P([FeP_1^t = f_p]) P(FeS_1^t) P([\lambda FeSP_1^t = 1] \mid [FeP_1^t = f_p]\ FeS_1^t)\ . \end{aligned}$$

In the summation over variable FeS_1^t, the term $P([\lambda FeSP_1^t = 1] \mid [FeP_1^t = f_p]\ [FeS_1^t = f_s])$ is always 0 except when $f_p = f_s$, so that the summation can be collapsed:

$$\begin{aligned} & P([FeP_1^t = f_p] \mid [\lambda FeSP_1^t = 1]) \\ & \quad \propto P([FeP_1^t = f_p]) P([FeS_1^t = f_p]) P([\lambda FeSP_1^t = 1] \mid [FeP_1^t = f_p]\ [FeS_1^t = f_p]) \\ & \quad \propto P([FeP_1^t = f_p]) P([FeS_1^t = f_p])\ . \end{aligned}$$

Therefore, $P([FeP_1^t = f_p] \mid [\lambda FeSP_1^t = 1])$ is not independent of $P(FeS_1^t)$. In other words, setting the coherence variable $\lambda FeSP_1^t$ to value 1 "closes" the Bayesian switch: contrary to the previous case, here, sensory information in $P(FeS_1^t)$ is combined with the distribution $P(FeP_1^t)$, and the combination operator is, mathematically, a product of the two probability distributions.

A technical precision can be raised here. Even though the definition above could suggest that a coherence variable "forces" the variables that it connects

to be equal, this is not so in effect. Indeed, it is true that a coherence variable imposes equality during inference, but this merely allows to "collapse summations" over the adjacent variables. This results in mathematical forms with products of probability distributions, such as, in our example, $P([FeP_1^t = f_p])P([FeS_1^t = f_p])$. In this expression, whereas it is true that the product is performed "assuming that variables have the same value", this does not imply any constraints on probability distributions $P(FeP_1^t)$ and $P(FeS_1^t)$. Indeed, these can "mostly agree", with their probability masses concentrated on the same values in their domain, or these can be "widely in conflict", with their probability masses on different portions of their domain, or any other situation in between. The mathematical machinery of coherence variables is agnostic to this and always results in a "fusion model" that is a product of distribution.

We now consider a slightly more complex example, in which coherence variable $\lambda FeSP_1^t$ would get controlled by an additional variable, A_1^t. The decomposition of the joint probability distribution $P(FeP_1^t\ FeS_1^t\ \lambda FeSP_1^t\ A_1^t)$ would become:

$$P(FeP_1^t\ FeS_1^t\ \lambda FeSP_1^t\ A_1^t) = P(FeP_1^t)P(FeS_1^t)P(A_1^t)P(\lambda FeSP_1^t \mid FeP_1^t\ FeS_1^t\ A_1^t)\ ,$$

with the term $P(\lambda FeSP_1^t \mid FeP_1^t\ FeS_1^t\ A_1^t)$ defined by:

$$\begin{aligned} &P([\lambda FeSP_1^t = 1] \mid [FeS_j^t = f_s]\ [FeP_j^t = f_p]\ [A_1^t = a]) \\ &= \begin{cases} 1 & \text{if } f_s = f_p \text{ and } a = 1 \\ 0 & \text{if } f_s \neq f_p \text{ and } a = 1 \\ 1/|Fe| & \text{if } a = 0. \end{cases} \end{aligned}$$

This is the same definition for this term as in the full model described above.

We now demonstrate that, with this definition of the model, the controlled coherence variable allows gradual control of information propagation in the model (Phénix, 2018). We consider, as above, computing:

$$\begin{aligned} &P([FeP_1^t = f_p] \mid [\lambda FeSP_1^t = 1]) \\ &\propto \sum_{FeS_1^t, A^t} P([\lambda FeSP_1^t = 1]\ [FeP_1^t = f_p]\ FeS_1^t\ A_1^t) \\ &\propto \sum_{FeS_1^t, A_1^t} P([FeP_1^t = f_p])P(FeS_1^t)P(A_1^t)P([\lambda FeSP_1^t = 1] \mid [FeP_1^t = f_p]\ FeS_1^t\ A_1^t)\ . \\ &\propto P([FeP_1^t = f_p]) \begin{pmatrix} P([A_1^t = 1]) \sum_{FeS_1^t} P(FeS_1^t)P([\lambda FeSP_1^t = 1] \mid [FeP_1^t = f_p]\ FeS_1^t\ [A_1^t = 1]) \\ +P([A_1^t = 0]) \sum_{FeS_1^t} P(FeS_1^t)P([\lambda FeSP_1^t = 1] \mid [FeP_1^t = f_p]\ FeS_1^t\ [A_1^t = 0]) \end{pmatrix} \\ &\propto P([FeP_1^t = f_p]) \left(P([A_1^t = 1])P([FeS_1^t = f_p]) + P([A_1^t = 0])\frac{1}{|Fe|} \right)\ . \end{aligned}$$

As in Equation (1), we note $P([A_1^t = 1]) = \alpha_{Fe}$, and rewrite this last result:

$$\begin{aligned} & P([FeP_1^t = f_p] \mid [\lambda FeSP_1^t = 1]) \\ & \quad \propto \quad P([FeP_1^t = f_p]) \left(\alpha_{Fe} P([FeS_1^t = f_p]) + (1 - \alpha_{Fe}) \frac{1}{|Fe|} \right) . \end{aligned}$$

To interpret this result, consider two extreme cases. First, when $\alpha_{Fe} = 1$, this result is identical to the simple case, and the two distributions over FeP_1^t and FeS_1^t are multiplied together: therefore, $\alpha_{Fe} = 1$ would correspond to fully closing the Bayesian switch. Second, when $\alpha_{Fe} = 0$, the distribution over FeP_1^t is multiplied with a uniform distribution, which leaves it unchanged (the uniform distribution is the neutral element for the product of probability distributions), so that the distribution over FeP_1^t is independent of the one over FeS_1^t: therefore, $\alpha_{Fe} = 0$ would correspond to fully opening the Bayesian switch. In the general case, however, α_{Fe} is neither 0 nor 1, which allows mixing the above two computations: the Bayesian switch is simultaneously "open" and "closed", in amounts controlled by α_{Fe}. This allows controlling, in a gradual manner, the amount of information propagated from the sensory distribution to the perceptual model. Therefore, the controlled coherence variable structure can be interpreted, not as a Bayesian switch, but as a Bayesian potentiometer (to pursue the electric analogy; a potentiometer allows gradual control of electric resistance, whereas a switch controls it in an all-or-nothing manner).

In the context of our model of speech decoding, variables $A_{1:15}^t$ are in charge of controlling which phone or syllable decoder receives sensory information; their distributions pilot all links between decoders and their respective sensory inputs. When the probability that the control variable A_i^t is True is 0, the corresponding decoder is not yet activated or already terminated; on the other hand, when the probability that the control variable A_i^t is True has a small, non-zero value, the corresponding decoder is currently activated, so that a small amount of sensory information is fed into the perceptual model.