Bayesian modeling of speech motor planning : variability, multisensory goals and perceptuo-motor interactions

Jean-François Patri

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **CIA – « Ingénierie de la Cognition, de l'interaction, de l'Apprentissage et de la création »**

Arrêté ministériel : 25 mai 2016

Présentée par

Jean-François PATRI

Thèse dirigée par **Pascal PERRIER**
et codirigée par **Julien DIARD** et par **Jean-Luc SCHWARTZ**

préparée au sein du **Laboratoire Grenoble Images Parole Signal Automatique (GIPSA-Lab)**
dans l'**École doctorale EDISCE – « Ingénierie pour la Santé, la Cognition et l'Environnement »**

Modélisation Bayésienne de la planification motrice de la parole

variabilité, buts multisensoriels et intéractions perceptuo-motrices

Bayesian modeling of speech motor planning

variability, multisensory goals and perceptuo-motor interactions

Thèse soutenue publiquement le **14 Juin 2018**
Devant le jury composé de :

Monsieur PASCAL PERRIER
PROFESSEUR, GRENOBLE INP, Directeur de thèse
Monsieur JOHN HOUDE
PROFESSEUR ASSOCIÉ, UNIVERSITÉ DE CALIFORNIE À SAN FRANCISCO, Rapporteur
Monsieur JACQUES DROULEZ
DIRECTEUR DE RECHERCHE ÉMÉRITE, CNRS DELEGATION PARIS, Rapporteur
Monsieur DAVID OSTRY
PROFESSEUR, UNIVERSITÉ MCGILL QUEBEC-CANADA, Examinateur
Monsieur JULIEN DIARD
CHARGÉ DE RECHERCHE, CNRS DÉLÉGATION ALPES, Co-Directeur de thèse

Monsieur EMMANUEL MAZER
DIRECTEUR DE RECHERCHE, INRIA, Président

Contents

List of Figures

Chapter 1

Acknowledgments - Remerciements - Agradecimientos

This acknowledgments section goes in many languages, but still, I feel that it is not enough to express all the gratitude to those that helped me and supported me in completing this thesis. I begin in English in order to address my gratitude to the four jury members – Emmanuel Mazer, Jacques Droulez, David Ostry and John Houde– for accepting to evaluate my work, for their time and involvement in reading the manuscript, and their effort in traveling from far to be present at the defense in Grenoble, or waking up early for the visio-conference due to the time difference with San Francisco. I sincerely thank all the interest and rigor of their questions and feedbacks.

Je bascule en français, afin de remercier encore et toujours tout l'investissement, le soutien, la générosité et la bienveillance de ces trois magnifiques personnes que j'ai eu la chance d'avoir comme encadrants. Pascal, Julien et Jean-Luc, j'ai vraiment beaucoup appris de vous au cours de ces années, tant sur le plan humain que scientifique. J'admire vraiment tout le temps, l'attention, la passion et le cœur que vous donnez et transmettez chaque jour à chacun d'entre nous. Pascal c'est surtout toi que j'ai le plus sollicité ces années, toujours t'arrangeant pour être disponible, la porte grande ouverte, la moustache souriante. J'admirerai toujours l'énergie et la motivation que tu dégages, tant dans la recherche qu'en dehors, toujours à vélo quel que soit le temps, même sous la pluie ou sous la neige. Force physique, force d'esprit et force de caractère, c'est sûr que ça doit faire peur quand ta moustache est sérieuse, mais moi j'ai toujours eu droit à une moustache souriante. Merci pour ta bienveillance, pour tout le temps que tu m'as accordé et toutes les choses que tu m'as apprises, sur le contrôle moteur que tu m'as fait découvrir avec passion, puis plus généralement sur la vie très souvent, avec tes histoires et tes opinions. Julien, toi aussi je t'ai sollicité sans trop me retenir, beaucoup par email étant donné les quelques rues à franchir pour aller dans ton bureau, mais combien de déplacements tu as dû faire pour des réunions régulières. Malgré ton emploi du temps souvent débordé ces dernières années (c'est ça le succès d'un encadrant investi !) tu as toujours été super réactif et disponible. Merci beaucoup pour cette présence, pour ton ton toujours décontracté, et surtout pour ta super pédagogie. Je suis arrivé dans cette thèse attiré par la modélisation Bayésienne de la cognition, et j'en repars ravi de l'avoir abordée sous l'angle de ton approche "algorithmique". Merci aussi pour toute la rigueur et la méthodologie que tu t'es acharné à me transmettre en ce qui concerne l'écriture. SPRI restera gravé en moi et me fera toujours penser à toi ! Tu resteras également ma référence en ce qui concerne le repérage des coquilles ou des espaces en trop dans un texte ! J'espère que ce texte ne t'agacera pas trop pour cela... Enfin, Jean-Luc c'est sûr qu'envers toi je me suis bien plus retenu de te solliciter, voyant le nombre de choses que tu dois gérer pour diriger ce projet ERC, plus toutes les autres choses que tu gères en plus et qui me dépassent. Mais quand même, malgré tout ça tu as toujours été présent quand j'en ai eu besoin, le sourire aux lèvres, la voix chantante et l'esprit aiguisé, pour me donner toujours les bons conseils et poser toujours les bonnes questions. Je ne te dirais jamais assez merci pour m'avoir fait confiance en

m'engageant comme doctorant, alors que tu m'as vu démissionner d'une autre thèse quelques années auparavant quand tu étais directeur de l'école doctorale [1]. Merci donc à vous trois, travailler auprès de vous a été un vrai plaisir, même pendant les six derniers mois un peu plus difficiles de rédaction et de soutenance. Vous resterez toujours des exemples pour moi. Merci encore et merci toujours !

Je continue toujours en français pour remercier toute la chaleur des interactions humaines au GIPSA-Lab. Tant du côté des chercheurs, du pôle ressources et du pôle technique, merci pour tous les services rendus, et pour ces beaux sourires qui allaient avec, c'est un détail, mais ça fait beaucoup ! En particulier, un merci très spécial au service informatique, que j'ai dû agacer plus d'une fois avec mes problèmes d'ordinateur, de carte son ou je ne sais quoi... En particulier Mikaël et Olivier pour le nombre de services de privilégié qu'ils m'ont rendus. Mikaël surtout, il faut dire, l'homme de la situation, se déplaçant toujours, prêtant sa langue à la science, et même apparaissant pile au bon moment en bas de chez moi pour me prêter un vélo le seul jour où je découvre le mien crevé, en retard pour aller faire passer un sujet d'expérience. Tu m'auras vraiment dépanné Mika, même avec les verres après le pot de soutenance ! Tu gères aussi fort que ce que tu serres les croûtes au pan d'escalade, vraiment Merci !!

Merci aussi pour tout le temps, les discussions, le support, l'aide, les conseils de toutes sortes, mais aussi les simples mots et sourires des chercheurs, doctorants et post doctorants du GIPSA. En particulier au cours de la réalisation de ces expériences avec l'EMA qui ont été un vrai casse-tête pour moi, merci à Thomas, Maeva et Nathalie qui m'ont tous donné un peu de leur temps à un moment où un autre avec mes questions et multiples problèmes (aussi Xavier et Sophie pour leur support technique avec mes cauchemars de faux contacts !)... Un grand merci à Christophe pour m'avoir toujours aidé et supporté avec patience lors de ces nombreuses tentatives de manipes pilotes où les problèmes ne manquaient jamais... Merci toujours pour ta bonne humeur, j'en aurai appris des choses de M. Savariaux ! Aussi, un énorme merci très spécial pour Pamela, qui est venue avec son organisation, son accent et son humour tout droit de Montréal et sans qui cette manipe n'aurait vraiment jamais abouti. L'efficacité et les fous rires ont vraiment eu un beau pic positif pendant ces mois où tu nous as nourris de ta bonne humeur et ton magnifique accent ! Tu as largement compensé les cauchemars et le découragement que j'ai pu avoir avec cette manipe avant que tu arrives ! Enfin, pour finir avec cette saga manipe, un énorme merci à Silvain aussi, pour sa disponibilité, sa patience et son expertise pour les analyses statistiques. Merci pour tout le surplus de boulot que je t'ai ajouté petit à petit, et surtout merci pour ton humour, discret, mais significativement génial ($p<0.00001$)!

Enfin, je ne peux pas faire le tour du GIPSA sans mentionner les complices de tous ces moments quotidiens au bureau. Les co-bureaux, Alexandre, Jonathan et Marie-Lou, puis Mélaine aussi pendant quelque temps, partageant ces innombrables énigmes de Marie-Lou au tableau et supportant malgré eux le niveau pourri de mes blagues. Aussi, les non moins proches co-cobureau juste à côté, Tiphaine, Thibault et Clémence le temps qu'elle a été là. Merci pour tous les rires, les escapades de grimpe et les footings à midi, les discussions sérieuses, scientifiques et personnelles, et celles moins sérieuses qui ne manquaient jamais non plus aussi... Merci d'avoir organisé ces beaux cadeaux pour la soutenance ! Vous avez trop géré ! C'est une sacrée chance d'être tombé parmi vous ! J'espère que nos chemins continueront de se croiser (ce ne serait pas la première fois, hein Tiphaine) !

Si une thèse c'est beaucoup de travail et d'interactions au labo, la contrepartie est quand

[1] Je profite ici pour faire un clin d'œil et remercier aussi Mirta Gordon, avec qui j'avais commencé cette autre thèse. C'est en particulier grâce à Mirta que j'ai pu reprendre mon parcours d'études, grâce au fait qu'elle a soutenu ma candidature au master de sciences cognitives trois ans après avoir démissionné auprès d'elle. Je me suis toujours dit que je te garderai une place dans mes remerciements !

même beaucoup de travail aussi mais avec beaucoup moins d'interactions en dehors. Ces lignes alors, toujours en français, pour tous les bons amis de l'hexagone, qui malgré l'asocial que j'ai été ces dernières années, sont restés et resterons toujours proches (enfin j'espère !); amis, presque famille, de grimpe beaucoup (malgré le peu de grimpe), de délires très souvent aussi, mais pas que. En particulier, je dois des remerciements spéciaux au puits d'informations, de discussions enrichissantes et de dépannages informatiques qui s'incarnent dans Le Mex. Combien d'infos précieuses tu m'as données toutes ces années ! Tout le monde devrait avoir un Mex pas loin de lui, Google s'en approche peut-être, mais il ne fait pas encore du bon pain et du bon toum comme toi ! Merci aussi aux autres amis de Grenoble et alentours, que je ne peux pas nommer tous par des raisons évidentes de place. En particulier celles et ceux d'escapades de grimpe, les soirs à la salle d'escalade, mais aussi dehors sous l'orage et à la frontale ! Puis ceux un peu plus éparpillés partout, un Blairi avec tête de Descartes en Ariège (et avec une chouette famille au Planious, plus son extension en Australie !), une bicyclette bleue à Briançon, un ardéchois à Thonon, un mouton fou à Valbonnais, son frère à Pont en Royans, puis Saint Marcellin, Annecy, Chambéry... Certains j'ai pu les voir un peu plus, d'autres pratiquement pas. J'espère me rattraper les mois qui restent ! Idem avec ceux qui sont déjà partis dehors, en Allemagne, en Angleterre, en Italie... Ja sama asia, mutta suomenkielinen, sirkus perheelle Helsingissä, että olen aina lähellä sydäntäni (I would be suprised if this actually means what it was intended to mean..., but well, a special thought to the great family of Circus Helsinki!).

Je finis la tournée en français pour ces deux beaux yeux bleus et cette magnifique longue tresse (puis tout le reste qui va avec qui est vraiment pas mal du tout non plus...) dont l'intelligence, la patience, la douceur et l'amour doivent sans doute venir d'une autre planète. C'est peut-être pour ça que j'ai parfois l'impression de te voir comme un animal exotique ? Tu es juste épatante ! Merci de continuer à m'accepter et me supporter depuis toutes ces années, malgré que je sois si souvent absorbé dans du travail... Des mercis je pourrais en remplir des pages, mais pour rester court et discret, je me restreins juste à deux très grands mercis : un énorme pour toutes le fois où tu t'es portée volontaire et enthousiaste pour participer comme cobaye pilote dans mes manipes, et un autre très particulier aussi pour ce joli pot de thèse que tu as si bien géré ! (et d'ailleurs merci aussi à Jérôme pour son aide si précieuse pour ça !). J'ajoute aussi un merci très spécial pour ta grande et chouette famille, qui m'a si chaleureusement adopté au cours de toutes ces années. Je vous dois quand même beaucoup d'excuses pour toute mon absence de ces dernières années, j'essaierai de me rattraper avant qu'on reparte pour de nouvelles aventures !

Y por fin termino en español, dirigiéndome a mis amigos y a mi gran y hermosa familia en Chile que extraño tanto. Recordarlos me llena siempre de cariño pero también a veces de nostalgia. Todos bien desparramados por Chile, en Santiago y sus alrededores, besotes a las primas y primos, tíos y tías, amigos y amigas, que a pesar del poco contacto guardo siempre con cariño en mis recuerdos. Yendo hacia la costa, besos y cariños también hacia Viña y la Serena, y pasando por Curacaví un abrazo fuerte a un entomólogo chiflado que me enseñó a escalar (¡y harto más!) y siguiendo hacia el sur, hacia Rupanco, más cariño para mucha linda gente. En especial por allá un Nachito del monte, en su pampa, con el que aún me veo haciendo portadas o tirándonos piqueros desde ese tablón al lago, mas su linda y gran familia, y un Lautarin payasín con risa contagiosa... y bueno, la lista de la tropa de lindos amigos y familia sería un poco larga, sin contar los que ya estan desparramados por el mundo, en Canadá, Australia, USA, y otros por aquí más cerquita, en Paris, Genève, Toulouse, Bordeaux, Angoulem, Bélgica, Escocia, Polonia... pero que igual de tan absorbido bien poco los he visto...

Los que si que se merecen una mención más que honrosa es la preciosa familia de Montpellier, que me adoptaron con tanto cariño hace ya 15 años cuando venía como pollito nuevo llegando

a Francia. Muchisisisimas gracias muy especiales por todo lo que me han dado y han hecho por mi todos estos años. Desde la primera vez que me trajeron en auto à Grenoble, hasta el pique que se pegaron para venir especialmente para la defensa. Jacqui (¡el mejor padrino!), Jerôme, Christophe y Lea, tenerlos ahí ese día fue lo mejor!! Y Carola, que aunque no pudiste venir, te sobre-que-requete pasaste levantándote tempranísimo para hacer esos deliciosos alfajorcitos que mandaste, junto con una linda tarjeta y una estampilla perfecta para la ocasión. Muchas gracias por todo ese cariño y atención, para ese día en particular, pero sobre todo a lo largo de todos estos años. ¡Los adoro!

Bueno, y que decir ahora para los que han seguido siempre tan cerca a pesar de los kilómetros y kilómetros hasta Chile. Mama, y Tata en las Vertientes, Papa, Quelly y Sophie en Osorno, Mémé en Santiago, Nicole, Negro y el Tokito en Rapa Nui, Thierry y Colette en constante movimiento... Siempre todos muy presentes, mandándome cariño y amor constante e incondicional, a pesar de lo desaparecido que puedo estar, a veces varias semanas sin dar noticias, absorbido siempre tan absorbido... Aunque sigo dando vueltas sin rumbo muy claro, es el tenerlos siempre en mí que me sigue dando paz y estabilidad. Seguido me parece escucharlos muy claramente, sus voces llamando "Jeaaan-Fraaan" como lo hacían para que baje de mi pieza en el árbol, o tu dulce voz Mama cantando las mañanitas para despertar. Yo parece que sigo en mi pieza en el árbol, y ustedes me llaman y me llaman para despertar y bajar. Me doy cuenta que no he cambiado mucho, sigo diciendo "al tiro" y sigo sin llegar. Pero ya llego..., ya llego....

Chapter 2

Introduction

1 An intriguing skill

Speech is probably the most intriguing and fascinating archetype of human skills. Somehow paradoxically, it is both ordinary and prodigious, controllable and uncontrollable, regular and variable. Indeed, speech is so ordinary that we are surprised when someone does not have it, or has it dysfluently, and we immediately attribute it to some kind of pathology. Speech is certainly ordinary because we produce it and perceive it with almost no effort. It is so spontaneous and we are so unaware of how we perform it, that it is often felt like a magical intangible substance just flowing straight through us, and yet it is felt as natural as having a shadow at our feet. Not only we produce and perceive speech without effort, but also we are often unable to avoid it. Unless covering our ears and shouting "la la la la", we just cannot avoid understanding speech (in our language, of course). Producing speech may seem easier to inhibit (although apparently not for everyone...), but it is actually only overt articulation that we manage to inhibit, inner speech remains often irrepressible in our mind.

Although ordinary, speech is actually a prodigious human ability. It is prodigious in how every normally developing child manages to master it without being explicitly taught. It is prodigious in how precisely, flexibly and effortlessly we perform this fast and complex stream of sounds with almost no errors, even though we achieve them under strong biomechanical constraints of organs, which evolved long ago for other unrelated functions (a feature referred to as "exaptation" in evolutionary biology). Indeed, the primary functions of our speech organs are eating and breathing, not speaking. Perhaps the evident situation in which we realize the prodigy of speech is when we travel to a place where people speak a language that we do not understand. Suddenly we are struck by how ordinary people produce complicated and incomprehensible sounds at an incredibly fast rate, and behave naturally, as if they would clearly understand each other in this annoyingly complex and noisy environment. Of course they do understand each other, and this highlights one of the most outstanding features of speech: precise messages are mediated by extremely variable and most often noisy sounds, and yet the intended meanings are almost instantly and effortlessly decoded by our mind.

This astonishing contrast between the variability of speech and the linguistic regularity underlying the meanings that we decode is indeed fascinating. It conveys both the beauty of speech and the complexity of its study, as illustrated by the difficulty of its processing and synthesis by speech technologies. Furthermore, speech is not just variable, it is variable in a variety of ways. There is of course variability between cultures, but also between regions within a same language. There is variability between genders, between ages, and there are even idiosyncrasies in the production and perception of speech for individuals belonging to the same social group. Adding up to these inter-subject differences, speech also features fundamental intra-subject variability that enables us to adapt speech to various contexts, situations and constraints. Furthermore, some of this variability plays a crucial functional role in communi-

cation, as it conveys linguistic and supra-linguistic features: we speak differently depending on our mood or depending to whom we speak. For instance, this enables us to mediate gravity or sarcasm. Another part of this intra-subject variability is fundamental in order to adapt speech to adverse conditions. For instance, we speak louder and over-articulate in noisy environments, we speak slower and with simpler sentences to children or foreign speakers, and we speak with different articulatory gestures while eating or holding a pen in our mouth. In brief, a large part of intra-subject variability is driven by features of the context. However, besides this contextual variability, a fundamental intrinsic variability, or token-to-token variability, remains: we never repeat exactly the same articulatory gesture twice, and each /a/ sound in "blah blah blah" is different from each other.

The intra-subject variability of speech emphasizes the complexity of processes that underlie its production. Not only are there plenty of ways to express a given thought, but also, can a given word and even a given phoneme be produced with several different articulatory patterns. How does our motor system select a specific action out of this abundance of choices?

2 Variability and the problem of choice

The starting point of this thesis is the question of how our motor system deals with variability at the planning stage of speech production, with a particular focus on the intrinsic token-to-token variability. Models of speech motor control have rarely considered this intrinsic source of variability, even though it is well known that it is this component that characterizes the naturalness of speech. In the general field of motor control, when this intrinsic variability is taken into account, it is generally attributed to noise in the neural processing of the articulatory execution pathway. However, while noise during execution certainly plays a role in movement variability, there are situations where it seems natural to assume that variability may also arise at a higher level of control and planning.

Imagine a situation where you are instructed to reach the perimeter of a square starting from its center. Computational approaches often formulate the selection of actions as rational choices that are deduced by solving equations that formalize the problem. However, in the present case, the task is under-constrained since there are an infinite number of possible points on the perimeter of the square that enable to successfully achieve the task. Computationally, these are called ill-posed problems, because they lead to equations that have multiple solutions from which it is impossible to deduce and decide rationally since they are all equally correct. The situation is similar to Buridan's ass, the equally hungry and thirsty donkey that dies of hunger and thirst when it is placed precisely midway between a stack of hay and a pail of water, incapable of making a choice.

To overcome this issue, a common approach is to consider additional constraints that may discard the excess of solutions and regularize the problem. For instance, optimal control approaches consider that regularities in the selection of actions arise from optimality principles. In the present reaching example, an optimality assumption may be to suppose that we select actions that minimize our displacements. Unfortunately there are two main shortcomings with such approaches. First, while optimality principles may certainly play a role in planning and executing movements, they cannot always ensure the uniqueness of optimal solutions. For instance in the present case, although the optimality assumption indeed drastically reduces the number of possible solutions, there still remain four closest (optimal) points from the starting position at the center of the square, and the situation remains undecidable. In order to fully regularize the problem, an additional constraint would be needed in order to break the symmetry of the task and eliminate the excess of solutions. The second main shortcoming of these

approaches is that whenever a single optimal solution is successfully obtained, the resulting planned action is inevitably stereotyped and trial-to-trial variability can only be attributed to execution noise or planning implementation approximations that induce deviations from optimality.

We certainly face hundreds of such ill-posed problem in our everyday life and of course even the donkey would manage to make a choice between the hay and the water. Our nervous system has evolved in order to act and perceive in a world full of uncertainties, constantly dealing with noisy and incomplete information and ill-posed problems are certainly part of these situations. Rationality does not imply getting stuck in such dilemmatic situations; it is rather the strict computational formulation of such problems that leads to their undecidability. Whenever sufficient information is not at hand, it is impossible to perform deductive reasoning. However, this does not prevent us from performing decisions. An intuitive view would consider that whenever we have multiple possible equivalent choices, we may well decide by selecting randomly among them, instead of getting stuck in a degenerate deductive decision process. However, how randomly? Uniformly randomly, with every choice having equal chance to be selected; or following some structure representing uncertainty, providing more chances to some choices and less to others?

Probability theory, in its subjectivist interpretation, provides a computational framework that enables to perform rational reasoning while dealing with uncertain or incomplete information. Combining this framework with a decision process that performs random sampling according to the computed probability distributions enables then to reproduce the variability of productions in the output space of the planning process.

3 Main goal and method

The main goal of this thesis is to apply probabilistic modeling in order to address the intrinsic and contextual sources of variability of speech production in an integrated framework. To do so, we postulate that the intrinsic, token-to-token variability of speech is more than the outcome of noise in the execution chain. We argue that variability is fundamentally involved at the representational level of speech motor planning (e.g., in the representations of motor goals), and that it characterizes the uncertainty resulting from the abundance of possible realizations of a given speech item.

We formalize this idea in a probabilistic modeling scheme, the Bayesian Programming framework, which, when applied to building models of cognition, becomes a framework for Bayesian cognitive modeling. In this thesis, we apply this modeling approach with three main contributions, addressing three main questions, which we now briefly introduce.

3.1 Intrinsic variability – Reformulation of GEPPETO

Firstly, we begin by presenting a reformulation in the Bayesian framework of an existing model of speech motor control, the GEPPETO model. For simplicity we call B-GEPPETO this Bayesian reformulation. GEPPETO is based on an optimal control scheme for the planning of phoneme sequences and accounts for fundamental features of contextual variability, such as coarticulation and prosodic features, including speech rate and stress. However, its optimal control approach prevents it from accounting for the intrinsic token-to-token variability in formal terms; an issue, as we have argued, that is also faced by other current models of speech motor control. By reformulating GEPPETO in Bayesian terms our goal is two-fold. First, we aim at formalizing the intrinsic variability that characterizes the production of speech items,

either isolated phonemes or sequences of phonemes. Second, we aim at incorporating the same theoretical ideas as in GEPPETO but formulating them in probabilistic terms. By this mean, we illustrate the greater power of expression of the Bayesian modeling approach, which enables to formulate the same general principles assumed to guide speech motor planning in GEPPETO, while maintaining uncertainty and variability at a representational level. In particular, this enables to maintain optimality without necessarily precluding the existence of variability at the planning stage of speech production.

Following this reformulation, we reexamine a particular feature of the B-GEPPETO, the involvement of a categorical perception module in its planning process, and compare it with an alternative planning architecture. We discuss the predictions of both approaches with respect to the structure of the resulting distributions of planned outputs.

3.2 Multisensory characterization of speech motor goals – Extension of GEPPETO

Secondly, a current shortcoming of GEPPETO is that it considers speech motor goals as being specified only in auditory terms. However, experimental findings suggest that speech motor goals may be characterized both in auditory and somatosensory terms and some current models have already included this feature. We propose an extension of GEPPETO in order to include both auditory and somatosensory specifications of speech motor goals. This involves some specific hypotheses concerning the existence of somatosensory characterizations of phonemes. We conceived and conducted two original experimental studies in order to evaluate the pertinence of these assumptions.

The fact that speech motor goals may be characterized in two different sensory modalities further suggests additional origins of both inter-subject and intra-subject variability. Indeed, the reliance on each sensory modality in the planning and control of speech gestures may vary from speaker to speaker and may also be modulated by each speaker according to the context. While a possible intra-subject sensory modulation does not appear to have been explored yet, experimental evidence already suggest individual sensory preferences in the control of speech. Since no modeling study has been proposed in the literature with respect to these phenomena, we use the model to formulate different possible interpretations for such sensory preferences.

3.3 Perceptuo-motor interactions in speech – Application of the framework to the interpretation of experimental data

Finally, an interesting feature of the Bayesian modeling framework is the possibility that it offers for integrating, in a single scheme, knowledge involved both in speech production and speech perception. This feature is particularly relevant for the study of the interactions existing between perceptual and motor processes in speech. In this regard, experimental studies have shown evidence of shifts in perceptual boundaries resulting from speech motor learning induced by perturbations of the auditory feedback.

However, we argue that the interpretation of these experimental findings is severely handicapped by a lack of precise hypotheses about possible underlying mechanisms. Thus, we further illustrate the pertinence of the present framework by applying it to evaluate some hypotheses concerning the motor and auditory updates that may result from speech motor learning, in the context of various assumptions about the involvement of auditory and somatosensory pathways in speech perception.

4 Outline

The remaining of this manuscript is structured as follows.

Chapter 3 reviews current models of speech production in light of the three main questions formulated previously: the question of variability, the involvement of auditory and articulatory characterizations of speech motor goals, and the interaction between perceptual and motor processes in speech. The aim of this chapter is to highlight the main differences between models of speech motor control and motivate the pertinence of GEPPETO as the foundation for this work.

Chapter 4 presents a precise description of the GEPPETO model in order to explicitly identify the essential features and hypotheses underlying its Bayesian reformulation and extensions presented in the rest of the document.

Part I details our first contribution about the study of intrinsic variability in a Bayesian modeling approach to speech motor planning, with two chapters. **Chapter 5** presents the Bayesian reformulation of GEPPETO step-by-step. We begin by introducing the Bayesian Programming approach with a preliminary example, then consider the case of isolated phonemes and finally consider the planning of phoneme sequences. We evaluate model performance and illustrate how it accounts both for intrinsic variability and contextual variability with coarticulation and speech rates effects. We further demonstrate how the Bayesian approach includes the optimal control approach of GEPPETO as a special case. This chapter contains material featured in two publications (Patri, Diard, & Perrier, 2015; Patri, Perrier, & Diard, 2016), and was presented in two conferences (Patri, Diard, Schwartz, & Perrier, 2015a, 2015b).

Chapter 6 reexamines the involvement of the categorical perception module in B-GEPPETO and compares it with an alternative planning architecture.

Part II concerns our second contribution about the inclusion of a multisensory characterization of speech motor goals, with three chapters. **Chapter 7** presents the multisensory extension of GEPPETO. We begin by defining somatosensory targets and evaluate model performance with simulations involving only somatosensory targets or both auditory and somatosensory targets. This chapter contains material featured in one publication (Patri, Diard, & Perrier, 2016).

Chapter 8 further develops the multisensory extension of GEPPETO in order to implement sensory preferences in speech motor planning. We describe two alternative ways for modulating the involvement of each sensory modality and evaluate their performance with simulations implementing auditory and somatosensory perturbations. This chapter contains material that was presented in conferences (Patri, Diard, & Perrier, 2017; Patri, Perrier, & Diard, 2017).

Chapter 9 presents two experimental studies conducted in order to evaluate two hypotheses formulated during the definition of somatosensory targets in Chapter 7. The first study aims at assessing differences in articulatory variability between utterances of vowels performed with and without auditory feedback. The second study explores whether subjects are able to correctly categorize vowels based only on proprioceptive information.

Part III deals with our third contribution, relative to the study of perceptuo-motor interactions in speech, with a single chapter, **Chapter 10**. We apply a simplified one-dimensional implementation of the multisensory Bayesian model in order to evaluate different hypotheses that may account for recent findings on perceptual shifts resulting from speech motor adaptation. This chapter reprints the entirety of one publication (Patri, Perrier, Schwartz, & Diard, 2018).

Chapter 11 summarizes the overall contributions of this thesis and suggests future directions to be explored.

Chapter 3

Challenges in speech motor control modeling

1 Introduction

This thesis presents a modeling work addressing the question of the variability in speech production, from the perspective of the definition of motor goals, the involvement of different sensory modalities, and the mechanisms required for the organization of control variables in order to reach these goals.

The aim of this chapter is to position our work in the context of the achievements of the main models of speech motor control. For that, we propose an overview of these models, which does not pretend to be comprehensive, but aims at providing a picture of the contributions of these models on the questions that we address in this work. Speech motor control models have been developed in the context of important theoretical debates between general motor control approaches. We believe that recalling the substance of these questions and debates is essential, in order to emphasize the fundamental distinctions between approaches. Therefore, we begin this chapter in Section 2 by summarizing the main context, questions and debates in the general field of motor control. In Section 3 we then focus on the particular case of speech motor control.

2 Context and fundamental debates in motor control

The aim of this first section is to review the main fundamental questions and debates distinguishing general motor control approaches. We begin by describing two main epistemological views concerning the origin of the observed regularities in behavior. We argue that distinctions between motor control approaches are best understood when they are considered in light of these two fundamental views. Within these two views there are further fundamental questions that distinguish motor control approaches. We then present an overview of these questions based on how the existing approaches describe the three main elements involved in the neural control of movements: the controller, the plant and the task.

2.1 Two main approaches to understand the neural control of movements: normative and physical approaches

When we perform intentional movements, we have the impression that we control them purposefully and that we can drive them with all the freedom allowed by the biomechanical constraints of our body. However, despite this freedom there are strong regularities in how we perform our actions. The science of motor control aims at understanding how living systems control their

movements, and in particular what are the origins of these underlying regularities. Current motor control approaches account for the regularities underlying the neural control of movement according to two main views.

2.1.1 Normative approaches

The first view fundamentally rests on the idea that the central nervous system is mostly omnipotent in how it can control body movements. In this view, movement regularities are not the result of physical (biomechanical or neural) constraints of the system, but essentially result from properties of the task that is performed. In other words, patterns or control strategies of movements are assumed to be fully specified at an abstract (computational) level characterizing the task, and are then implemented through the effector system that does not alter them in any crucial way. Since under this approach properties of the task dictate and physical implementation follows, the computational approach is termed as "normative". The following quote characterizes well this view (Haith & Krakauer, 2013, p.7-8):

> The sheer flexibility of the motor system makes it seem unlikely that underlying mechanisms place a significant constraint on the kinds of movement that can be generated. Instead, it seems that regularities in behavior are mostly dictated by features of the task at hand rather than by features of the underlying implementational mechanism. A normative modeling approach seeks to explain behavior by first understanding the precise computational problem that the brain faces, and then asking what, theoretically, is the best possible way to solve it (akin to Marr's computational level of analysis). [...]
> The normative point of view effectively assumes that the underlying neural mechanisms have omnipotent capacity. Consequently, aspects of the task itself, rather than the underlying mechanisms responsible for implementing the solution, are what primarily dictate our patterns of behavior.

2.1.2 Physical approaches

The omnipotent assumption of normative approaches is certainly strong. As acknowledged by Shadmehr and Mussa Ivaldi (Shadmehr & Mussa-Ivaldi, 2012, p.6):

> We are not going to be asking about how neurons in the brain might actually perform the computations that are implied by our mathematics — our approach is going to be fraught with danger. After all, it is quite possible that we cannot know why an organism does something until we know how it was built. That is, we may not be able to successfully theorize about the regularity in the way that our brain moves our body until we are much farther along in understanding the basic facts regarding the biology of our brain and its evolutionary history.

Physical approaches are opposed to the normative view in that they consider that some physical and biological constraints of the system are fundamental for understanding the regularities of its behavior. In this view the central nervous system is not omnipotent, but control movements by exploiting specific biomechanical and neurophysiological properties of the controlled body (Balasubramaniam & Feldman, 2004; Latash, 2010).

2.1.3 Illustrating normative and physical approaches with an example

The distinction between normative and physical approaches can be best illustrated with a simple example. Imagine that you are given a pendulum composed of a mass M attached to a point with a rod of a length L. The external gravitational force drives the pendulum to swing in a specific way, with a period that is proportional to the square root of the rod's length. Your task is to make the pendulum swing at a different period, and for this your only device is a motor that could act either as a rotor, applying torques to the rod, or as a lifting-lowering device, modifying the length of the rod.

The task can be performed in two ways that illustrate the distinction between normative and physical approaches of motor control. Following the first approach you would use the rotor in order to apply torques to the rod and precisely control the movements of the mass. This approach is omnipotent, since you have complete freedom on the kind of movements that you can impose to the mass. In particular you can perform the task irrespective of whether the gravitational force is present or not, you would only need to adapt the required torques in the presence or absence of gravity. However under this approach the task appears to be strongly under specified, since the intended period could be performed with different patterns of movements and with different lengths of the rod. Because of this abundance of freedom, you would need to select, arbitrarily, or based on some preference, a specific pattern of movement out of all the available ones.

Following the second approach, you would perform the task by exploiting an intrinsic property of the system: its physical dynamics. In order to change the swinging period of the pendulum, you would only need to shorten or lengthen the rod. This approach is not omnipotent since it critically rests and exploits the intrinsic dynamics of the pendulum system by modifying its length parameter. In particular, this approach does not work in the absence of gravity: the physical implementation of the problem is therefore crucial.

Note that contrary to the normative approach, the particular swinging movement of the pendulum in the physical approach results as a direct consequence of its intrinsic physical dynamics. However, the normative approach could also "emulate" the same sinusoidal displacement of the mass as obtained by the physical approach. For this, the normative approach would only need to impose the same (virtual) dynamic to the mass. The crucial difference is that in the normative approach the actual length of the rod and the external gravitational force are secondary implementational parameters, which in this specific case would affect the torques applied by the rotor, but not at all the patterns of movement, while in the physical approach they are fundamental features determining the patterns of movement.

In summary, physical approaches address control problems in a more physically constrained view than normative approaches. Physical features of the system are crucial for physical approaches, while they are secondary and considered only at an implementational stage for normative approaches. Physical approaches expect that behavioral regularities result from how biological systems exploit physical constraints to their advantage, whereas normative approaches expect that behavioral regularities result from computational principles at the abstract level of the task.

While opposed in their views, both approaches are certainly appealing and complementary. As pointed out by Haith and Krakauer (Haith & Krakauer, 2013, p.8): "Mechanistic and normative approaches are far from mutually exclusive endeavors – breakthroughs in normative models of behavior often inspire and help guide mechanistic models. A deeper mechanistic understanding can help to constrain normative models".

2.2 The controller, the plant and the task

Motor control approaches are generally formulated in terms of how the controller (the central nervous system) organizes the time variations of the control variables that drive movements of the controlled the plant (the body parts), in order to perform a particular task. Motor control models can be therefore distinguished according to three main questions: (1) how do models describe the task, in terms of motor goals, a question which is essentially related to what models assume to be the nature of the controlled variables; (2) how do models describe the plant, a question which is essentially related to what models assume to be the nature of the control variables and their consequences on the controlled variables; and (3) how do model specify the controller, in terms of control strategy and control architecture, in order to solve the control problem.

2.2.1 The task – nature of motor goals and controlled variables?

The first question that needs to be addressed in order to describe how movements are controlled, is how the controller represents the task that is intended to be achieved. Answering this question requires translating the abstract formulation of the task ("reach the target", "say vowel /a/") into a concrete specification of the goal that is intended to be achieved, and the variables that are controlled for it. For instance, in the case of a reaching task where the task would be formulated as "reach the target with your finger" the goal is to position the finger in a particular location in space. How is this location represented in the controller? Is it through visual information, proprioceptive information, or both? And what are the variables that are controlled in order to achieve this position? Is it finger position, or joint angles (Ghez, Scheidt, & Heijink, 2007)?

This question is particularly relevant in speech, where the task is to communicate some abstract linguistic information. As nicely summarized by Grimme and colleagues in a conceptual review comparing limb versus speech motor control (Grimme, Fuchs, Perrier, & Schöner, 2011, p.9):

> Contrary to object-oriented reaching or grasping movements, the objectives of speech production are defined relative to linguistic information and communicative intentions of the speaker, but not relative to the directly accessible physical world. For most communicative intentions, the physical characteristics of speech production have no meaning by themselves but make sense only in relation to the listener's perception. As a result, the physical correlates of speech production can be highly variable without invalidating the perception of speech.

So, how should the abstract linguistic goals of speech be represented in concrete terms? Do they correspond to auditory goals, articulatory goals, or both? Are these goals static, or are they fully specified in the temporal domain?

2.2.2 The plant – nature and impact of control variables?

The plant corresponds to the effector system, the part of the body that is controlled in order to achieve the intended task. Where is the boundary between the controller and the controlled part? What is the nature of the control variables sent by the controller in order to drive movements of the plant and which are their impact on the controlled variables? The nervous system is generally divided into Central Nervous System (brain, brainstem and spinal cord) and Peripheral Nervous System, which includes efferent and afferent neurons from the spinal cord (or the brainstem) to the muscles. Some reflex movements, such as those induced by the

muscle stretch reflex, are essentially induced by the low level interface between the peripheral nervous system and the spinal cord (or brainstem). Are such reflex mechanisms part of the plant or are there part of the controller?

This question is well-illustrated in the context of normative approaches. Indeed, for these approaches depending on where boundaries are put between the plant and the controller, different levels can be considered in the control problem, involving different control variables and consequently different relations between control and controlled variables.

At a first level the control problem does not take into account the effector system that performs the task. For instance, in the reaching example, the task could be defined as reaching a particular position in a 2-dimensional horizontal surface with a pointer. A solution to the control problem at this level would formulate a computational principle or strategy that would enable to account for the formation of the trajectory and velocity profile of the pointer in order to reach the intended target. This first level could be said to be purely computational, in the sense of Marr (1982), since the control problem does not take into account any representation of the effector system that performs the task.

At a second level of description, the effector system starts being taken into account, but in pure kinematic terms. Assuming a simplified arm with three joints - shoulder, elbow and wrist - this level would describe the arm in terms of the three corresponding joint angles in the 2-dimensional plane of the task. A solution to the control problem at this level would further account for the time evolution of each joint angle such that they drive movements of the pointer according to the control strategy specified in the previous level. As for the first level, this is a purely kinematic description of the control problem, but in the representational space of the effector instead of the computational space of the task.

A third level of description would take into account the physical dynamics of the effector by considering the forces – or in the present example joint torques- that would need to be produced, in order to change joint angle configurations. A solution to the control problem at this level would account for the time evolution of joint torques responsible for the temporal evolution of joint angles as specified in the previous level. The control problem at this level would involve knowledge about the dynamical parameters of the system (in particular the distribution of its mass), as well as external forces applied to the arm (for instance gravity or friction).

A fourth level of description would further consider what underlies the generation of the forces (or torques) specified in the previous level. In a robotic arm, torques are directly generated by motors, but in a biological arm torques result from several agonist and antagonist muscles surrounding each joint. Therefore, a solution to the control problem at this level would need to account for the time evolution of each of these individual muscle forces, in order to generate the overall forces or torques that drive movements of the arm.

A fifth level of description is then concerned with what determines the forces generated by each individual muscle. These forces can be decomposed into passive and active components; passive forces result from passive biomechanical properties of muscle, tissues and tendons, which in a rough approximation behave as linear springs; active muscle forces are generated by contraction of muscle fibers, triggered by alpha motor neurons located in the spinal cord and brainstem. Only this active component of muscle force can be actively controlled through neural signals to the muscle. Hence, a solution to the control problem at this level would account for the time evolution of alpha motor neurons activity, in order to correctly activate muscles and drive intended movements of the arm.

A sixth level of description would then ask what specifies the activation of alpha motor neurons. Alpha motor neurons are known to receive afferent inputs both from proprioceptive neurons (in particular muscle spindles) and efferent inputs from upper motor neurons located in

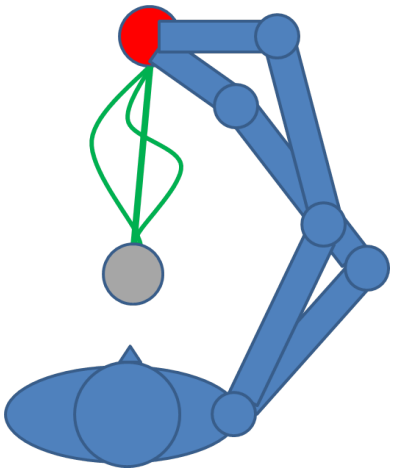

Figure 3.1: Trajectory formation and motor equivalence: *two aspect of the degrees of freedom problem in motor control illustrated with an arm reaching task.*

the motor cortex. Proprioceptive neurons are sensory neurons that provide information about the relative positions of body parts and their variations over time. They are mainly sensitive to deformations, providing information about muscle length (muscle spindles) and velocity (Golgi tendon organ). In particular, muscle spindles contribute to trigger activity in alpha motor neurons when a muscle is stretched. This results in an unintentional muscle contraction: the muscle stretch reflex. Should this reflex be considered as part of the controller? Or should it be part of the plant and cancelled out by the controller in order to perform the intended patterns of movements?

Within a normative approach, each of these levels adds further computational steps to the control problem. However, these problems are seen as additional precisions to be included in order to achieve a fully physical description of the plant that can be integrated in the computation in order to cancel its influence on movement, and they are often considered to be reasonably neglected as a first approximation. Within a physical approach these levels are fundamental, since they contain physical and biological features that influence movements, and on which the specification of the control strategy critically needs to rest.

2.2.3 The controller: Control strategy and architecture

The controller includes all the processes that plan and organize the time evolution of control variables, in order to drive movements of the plant toward the intended goals. Motor control approaches differ at this stage in the computational principles that they assume to be involved in these processes. The specification of the controller involves two main related questions: (1) what control strategy is assumed, in order to address the fact that in humans the motor system has degrees of freedom in excess, enabling different strategies to reach the same goals; (2) what control architecture is considered to implement the control strategy.

1- Control strategy: addressing the degree of freedom problem At almost every level in the description of the control problem there are more degrees of freedom to achieve the task than constraints specifying it. This is known as the degrees of freedom problem, redundancy problem or motor equivalence problem in motor control (Bernstein, 1967; Turvey, Fitch, & Tuller, 1982). It implies that at almost every level there are multiple alternatives leading to the same output. In the case of the arm reaching problem, at the first level there

are an infinite number of possible trajectories and velocity profiles of the pointer that end-up reaching the intended target (green curves in Fig 3.1). At the second level there are multiple alternative geometrical configurations of the effector that result in the same position of the pointer (Fig 3.1). At the fourth level there are multiple configurations of individual muscle forces producing the same torque or total force, and so on. How does the controller select a specific action out of all this abundance of possibilities?

We begin by focusing on the degrees of freedom problem at the level of trajectory formation. Different approaches have been proposed in this respect that can be distinguished according to the level at which they define their control strategy.

Trajectory formation at the level of the task The problem of trajectory formation can be directly addressed at the level of the task by providing a strategy that accounts for the performed trajectory. Several approaches have been suggested in this spirit, which can be grouped with respect to whether they separate or not planning and execution.

a) Planning followed by execution

The first possibility is to provide a principle for the selection of a particular trajectory prior to the execution of movements, which is then followed during the execution of movements. Possible principles could be ad-hoc or influenced by factors external to the task as in the case of some idiosynchratic choices, or based on optimality principles (for instance, minimizing displacement).

b) Planning during execution

The second possibility is to consider that trajectories are not selected in advance, but emerge from a particular action plan that enables to select movements at each time step. In the case of goal oriented movements, one possible approach is to provide an action plan in which the target defines a point attractor that guide movements towards it, whatever the initial position of the cursor or possible perturbations during execution.

Note that approaches that solve trajectory formation at the level of the task are necessarily normative, since their control strategy at this level does not involve any feature of the plant.

Trajectory formation in a proximal space The second family of approaches considers that the controller does not specify movements at the level of the task, but in a space closer to the control variables, a "proximal" space. For instance, in the arm reaching task, the controller could select one geometrical configuration of the arm (for example the joint angles space) for which the end-point coincides with the position of the target and then specify movements in this geometrical space (for instance as a linear interpolation between initial and target joint angles configurations (Rosenbaum, Loukopoulos, Meulenbroek, Vaughan, & Engelbrecht, 1995)).

Approaches can also address trajectory formation directly at the level of control variables. In this line, a class of optimal control approaches directly derives the patterning of control variables by optimizing some performance criterion directly in the space of the control variables, assuming knowledge of the physical dynamics of the system and its response to control variables. These approaches differ in whether they separate planning from execution (Harris & Wolpert, 1998), or perform motor planning during execution (Todorov & Jordan, 2002). Such optimality based approaches are currently the most often cited in the literature, and have accounted for a large number of behavioral observations (Todorov, 2004).

Another approach that addresses trajectory formation at the level of control variables is the Equilibrium Point Hypothesis. In this approach the plant includes the low level muscle stretch reflex, which gives to the plant stability and equifinality towards equilibrium configurations. Movements are triggered by changing the neurophysiological control variables (λ parameters)

characterizing the equilibrium configurations of the plant and physical trajectories toward new equilibrium state emerge from the interaction between the biomechanical features of the plant and the shift in λ parameters. In order to reach a given target, the controller then first need to identify a configuration of λ parameters that achieves the intended target, and then shift the actual λ parameters toward their target values. In this view, the trajectory in the task space is shaped by the specification of the targets and the dynamical behavior of the physical plant.

Note that while approaches that addressed the trajectory formation problem in the space of the task are per essence normative approaches, assuming the capacity of the controller to integrate all the functional complexity of the plant and to cancel its effect on the movement on the task space, approaches that address the problem in a proximal space are compatible with a somewhat normative and a somewhat physical approach. However different levels of "normativity" and "physicality" need to be considered. For instance, when the trajectory formation problem is treated in the space of the joint angles, it is assumed that the controller can fully manage the control of these angles, and is able to cancel the impact of biomechanical and dynamical factors on these angles. In that sense it is a normative approach in the treatment of the relation between control variables and joint angles. However, interpolating between two target joint angle configurations will have some impacts on the trajectory in the task space, which are not controlled and are just the consequence of the relation between joint angles and pointer position embedded in the properties of the arm. Optimal control approaches also include knowledge about the physics of the plant and its response to control variables, and the optimality criterions often include features of the system, as signal dependent noise of neural pathways (Harris & Wolpert, 1998; Todorov & Jordan, 2002). However, these approaches essentially rely on the idea that the controller perfectly knows the functional relation between control variables and the space in which the optimization operates in order to cancel its impact on the criterion to be optimized, and neglecting the nature of the biological phenomena involved in this functional relation such as muscle force generation principles or low level reflex mechanisms. At the opposite the Equilibrium point hypothesis relies in a greater level of physicality since the trajectory in the task space is largely influenced by biomechanical properties and low level reflex mechanisms of the plant.

Redundancy in trajectory formation problems Approaches that address trajectory formation in the space of the task face the degrees of freedom problem at all more proximal levels. Indeed, once a trajectory or an action plan is specified in the space of the task, this solution must be translated into the more proximal spaces in order to specify the control variables that achieve the intended pattern of behavior. For instance, for a given displacement of the pointer in the three joint arm example, the controller needs to specify the corresponding changes in joint angle, the corresponding torques that achieve these changes in joint angles, and so forth. Yet, the same displacement of the pointer can be performed by several changes in joint angles, and the same torques can be achieved by different configurations of individual muscle forces. Mapping movements in task space to movements in the articulatory space of the effector (known as inverse kinematics) or and to torques or forces (inverse dynamics) is therefore ill defined and further approaches need to be defined in order to solve redundancy at these subsequent levels.

Approaches that address trajectory formation at a proximal level only need to solve redundancy at a reduced number of levels, those separating the control variables and the proximal space in which trajectory formation principles are applied.

2- The controller Architecture A second fundamental question raised by the specification of the controller is how sensory feedback is assumed to be involved in the control strategy. The different control strategies described previously can be implemented by different control architecture according to how sensory feedback is taken into account. We briefly review three types of control architectures corresponding to whether they involve no feedback, direct sensory feedback, or internal feedback.

i) Feedforward control architectures – No feedback Feedforward control architectures implement control strategies that specify the whole temporal pattern of control variables in advance, prior to execution. These temporal patterns can be obtained from control strategies that address trajectory formation at the level of the task or at proximal levels. The fundamental feature of feedforward control is that movements are executed as a recorded tape without involvement of sensory feedback. This feature is both the main advantage and the major drawback of this control architecture. Its advantage is that it enables to account for how very fast movements (e.g. eye saccades, fast speech movement) are performed, since in such cases sensory feedback has no time to inform about the current state of the plant before the end of movement. Its drawback is that no error correction is possible during execution. The patterning of control variables needs to be extremely precise and any noise or perturbation can critically impact execution.

ii) Feedback control architectures Feedback control architectures implement control strategies that specify control variables during execution, based on the current state of the plant. Two main approaches have been proposed in order to account for how the current state of the plant is estimated, through sensory feedback or internal feedback.

Sensory Feedback: the only source of information that enables humans to access the real state of their body is sensory information. Sensory information can be integrated from various modalities (the most pertinent for motor control being proprioceptive, tactile, visual and auditory modalities) and at various hierarchical levels, from low level reflexes in the spinal cord or the brainstem to high level sensory integration in cortical sensory areas.

Advantages and drawbacks of feedback control architectures that rely only on sensory feedback are mirror images of feedforward architectures. Their main advantage is that they can cope with perturbations and adjust control variables during execution. However, sensory feedback is both delayed and noisy (in particular for high level sensory feedback which can take around 100 ms to be integrated, for instance in speech (Lechner, 1979)). When these features of biological systems are taken into account, pure sensory feedback control architectures become unstable and unreliable.

Internal Feedback: To overcome the issues faced by pure sensory feedback approaches, internal feedback approaches (such as the one implemented in State Feedback Control) consider that the state of the plant is estimated both by internal predictions, and sensory information. Internal predictions are assumed to be obtained from internal forward models in the brain, which enable to internally simulate the plant and predict its current state based on a copy of the control variables that are sent to the plant (generally termed as "efference copy"). This enables improving state estimation, by combining predictions and sensations, as well as coping with delays in neural pathways by taking them explicitly into account. The main advantage of the state feedback control architecture is that it combines both advantages and avoids both drawbacks of pure feedforward and pure sensory feedback control architectures: fast movements can be executed reliably by resting mainly on internal estimations, which are assumed to be

fast, and sensory information can be used in order to adjust control variables during execution for correcting errors and cope with perturbations.

iii) Hybrids Other hybrids control architectures have been defined that enable to avoid both drawbacks of pure feedforward and pure sensory feedback control architectures. Hybrid architectures essentially combine both feedback and feedforward components which can be organized both in series and in parallel. Speech motor control models provide two examples of hybrid control architectures. The DIVA model combines feedforward and feedback control architectures in parallel. The GEPPETO model, based on the Equilibrium point hypothesis, can be seen as a pure feedforward controller that feeds a low level (muscle stretch reflex) feedback controller at the level of the plant.

3 Questions and approaches in speech motor control

Models of speech motor control have been developed in the general context of questions and debates presented in the previous section. The aim of the current section is to highlight the main differences between models of speech motor control in the light of this general context. We consider three main models of speech motor control: Task Dynamics, DIVA, and GEPPETO. While this choice is not exhaustive, we focus on Task Dynamics and DIVA since they have been the most influential models of speech motor control for the last 20 years. We also include the GEPPETO model since it has been developed in our lab and is at the chore of the present work.

3.1 Overview of current models of speech motor control

We begin by presenting an overview of the three considered models by structuring them according to the two of the three main components that we discussed in Section 2.2, the task and the controller. We leave the question of the plant for the discussion.

3.1.1 Task Dynamics

The Task Dynamics model (Saltzman, 1986, 1991), in conjunction with the Articulatory Phonology (Browman & Goldstein, 1992) has been for long the only model of speech production that has integrated concepts coming from the field of motor control research (Bernstein, 1967; Turvey, 1978), to account for the link between the phonological level and speech movements. It has certainly been the most influential and debated model in the 80s and early 90s, and has put at the core of speech production research the notions of dynamics and gestures.

The speech task in the Task Dynamics model is considered to be a sequence of elementary gestures. Each of these elementary gestures is characterized in terms of vocal tract parameters specifying constriction locations and degrees. The sequence and timing of these gestures are specified in terms of gestural scores in the context of the Articulatory Phonology (Browman & Goldstein, 1992), which associates to phonemes units movements oriented toward vocal tract configurations rather than acoustic features, as proposed by Halle and Chomsky (1968). The Task Dynamics model has been used to control an articulatory model of the vocal tract (Mermelstein, 1973; Rubin et al., 1996).

The control strategy The Task Dynamics model addresses trajectory formation in the task space, with a control strategy that specifies actions during execution by considering the discrete constriction targets as dynamical attractors. These attractors correspond to damped second-order linear systems (a virtual mass-spring model) in the space of the task. The name of Task Dynamics comes from this idea that the dynamics of intended gestures fundamentally originates at the level of the task space. The precise time series of the discrete vocal tract targets (in terms of constriction locations and degrees), are considered as gestural scores specified by a higher level planning module and tightly linked with the phonological units in the context of the Articulatory Phonology (Saltzman & Munhall, 1989). Proposals have been made at the level of the syllable to relate the phonological structure with the timing of the gestural scores (Browman & Goldstein, 1988).

Consequently, intended movements in the task space need to be further translated into the space of articulatory control variables of the articulatory model, in order to perform actual movements of the plant. Translating intended movements in task space to the required time evolution of articulatory parameters requires an inverse kinematic step that is usually performed with a pseudo inverse of the Jacobian matrix (a linear approximation of the mapping relating infinitesimal displacements in articulatory space to displacements in task space).

The control architecture The control strategy in Task Dynamics dictates movements of the effector system by comparing the current state in the task space to the intended target configuration. This crucially involves precisely knowing the current state of the effector system at any time step, which is assumed to be possible through an ideal feedback observer based on somatosensory information. This makes Task Dynamics a purely feedback control architecture involving only somatosensory feedback. Whether this sensory feedback is integrated in low level structures (spinal cord-brain stem) or higher level structures in the Central Nervous system is not explicitly clarified.

Task Dynamics and State Feedback Control Recently, the control architecture of the Task Dynamics model has been improved with a State Feedback Control approach (Houde & Nagarajan, 2011; Ramanarayanan, Parrell, Goldstein, Nagarajan, & Houde, 2016). This development intends to overcome the limitation of the ideal observer in Task Dynamics, which assumes that states of the plant would be instantly and perfectly accessible continuously in time. This critical assumption is certainly false, since states of the effectors can only be accessible through sensory feedback, which are both noisy and delayed (Houde & Nagarajan, 2011). Another limitation of the Task Dynamics model is that it does not include the fact that the state of the vocal tract is not only accessible via somatosensory feedback, but also via the auditory feedback. This concept has also been introduced in the application of State Feedback Control to the model.

State Feedback Control proposes to overcome the instability that would result from including noise and delays in sensory feedback in a purely feedback control architecture. According to State Feedback Control, the state of the effector system would be accessible through a running time estimate obtained by combining both sensory feedback and internal predictions with an internal model through a Kalman filter approach. The new control architecture proposed by Ramanarayanan et al. (2016) then replaces the real plant state, assumed to be ideally observed in Task Dynamics, with an estimated plant state obtained with the State Feedback Control approach.

While Ramanarayanan et al. (2016) only considered auditory feedback in their implementation, in principle the State Feedback Control approach enables to involve and combine both

auditory and somatosensory feedback for the estimate of the state of the plant (Parrell, Ramanarayanan, Nagarajan, & Houde, 2018).

3.1.2 DIVA

The DIVA model is today the reference model of speech production, as it proposes a description going from the brain (and the cerebellum) to speech acoustics of how speech production is planned and controlled. It includes thus a comprehensive framework for analyzing and interpreting brain processes and regions involved in speech motor control. In particular, it is the only model that has been applied to qualitatively reproduce neural activity during speech production tasks.

The speech task DIVA first considered speech as a sequence of discrete phonemic motor goal regions. These regions were initially characterized in terms of acceptable ranges of vocal tract geometry (Guenther, 1995) for each phonemic unit, and were then reconsidered as auditory regions in auditory terms (Guenther, Hampson, & Johnson, 1998). The most recent version of DIVA considers both auditory and articulatory targets, but the discrete description of phonemic regions was abandoned and replaced with a time-dependent description of target regions corresponding to syllabic units which can be seen as "trajectory tunnels" with permissive upper and lower bounds.

The control strategy The control strategy in DIVA has some similarities with the one of the Task Dynamics model in that it also addresses trajectory formation at the level of the task and its control strategy also specifies actions during execution according to the current state of the plant and its distance towards the intended targets. However, the particular law of the control strategy differ between task dynamics and DIVA, since DIVA specifies velocities of displacement that are proportional to the direction and remaining distance from the current state of the plant towards the intended target region in task space. This control strategy gives its name to DIVA since it drives articulatory movements by mapping "Directions Into Velocities of Articulators".

DIVA also rests on a geometrical description of the plant (the MAEDA model (Maeda, 1990)) and therefore control variables are also identified with articulatory parameters of the geometrical model. Translating directions in the task space to velocities of these articulatory parameters further involves an ill-posed inverse kinematic problem, as in Task Dynamics (though with respect to different spaces). This inverse kinematic step has also been addressed with a pseudo inverse approach in order to invert the Jacobian matrix relating displacements in articulatory space to changes in auditory space.

The original description of DIVA in terms of discrete target units enabled it to account for trajectory formation and to propose a way to account for some effects associated with speed rate variation. However, in its current version the time dependent characterization of targets strongly constrains the resulting trajectory and speed rate effects can only be imposed by the controller.

The control architecture The control architecture in DIVA combines a feedback controller in parallel with a feedforward controller. As the current version of DIVA includes both auditory and articulatory targets, the feedback module combines the errors resulting from auditory and somatosensory pathways. The actual control signal hence result from the combination of two sensory feedback errors and the feedforward component. In the beginning of development, DIVA

considers that speech production is initially driven only by the feedback controller. During development, the feedforward module would then progressively learn, from the feedback controller, the control patterns involved in frequent speech sequences. Hence, after development, speech movements would be mainly driven by feedforward control, and feedback adjustments would be involved only in the case of production errors, or errors induced by external perturbations.

3.1.3 GEPPETO

The GEPPETO model has been designed in the middle of the 90's in order to implement the following fundamental hypotheses in a motor control model : (1) speech production is driven toward discrete physical targets that are related to phonemes; (2) articulatory trajectories over time and, hence, acoustical time variations are not fully controlled, but results from the specification of the targets and of the physical characteristics of the peripheral speech production system, (3) the selection of the motor strategies adapted to the production of speech sequences results from optimal planning taking the phonological structure into account. As compared to the Task Dynamics and the DIVA models, the GEPPETO model is then different in that it includes a biomechanical description of the orofacial system and gives optimal control a key-role in articulatory coordination. An exhaustive description of GEPPETO is proposed in Chapter 4. Hence, the description below is rapid and only aims at positioning the model with respect to the Task Dymanics and the DIVA models in the context of the key-issues that we address in this section.

The speech task GEPPETO considers the speech task as a sequence of discrete targets related to phonemes and essentially characterized in the acoustic domain. For GEPPETO the primary goal of speech gestures is thus auditory. Phonological units are described in GEPPETO by ellipsoid regions in 3-dimensional formant space, corresponding to the range of variability observed in production tasks.

The control strategy Contrary to DIVA and the Task Dynamics models, GEPPETO does not address trajectory formation at the level of the task, but directly at the level of control variables. GEPPETO rests on a biomechanical model of the plant in which muscle forces are assumed to be generated with respect to the Equilibrium point Hypothesis (Feldman, 1986). The equilibrium point hypothesis considers that the control of movements exploits a fundamental biological property of the effector system: the low level muscle stretch reflex. The muscle stretch reflex acts as a low level feedback controller that generates active force in a muscle whenever this muscle is stretched with respect to a reference length. The Equilibrium Point Hypothesis considers that the high level controller does not cancel out but exploits these low level reflex mechanisms and drives movements by modifying their reference lengths parameters (λ parameters). Movements are thus instantiated by shifting the values of these reference λ parameters and trajectories emerge from the interaction between these λ shifts, as specified by the high level controller, and the dynamical response of the effector with its low level reflex loops.

A particular interest of this approach is that it avoids the involvement of inverse kinematics and inverse dynamics which the authors consider to be computationally intractable problems when all the biophysical complexities of the system are taken into account. Instead, the problem faced by the high level controller is the appropriate shift of λ control variables in order to achieve the intended targets of the task. GEPPETO assumes that these shifts are simple linear transitions between specific λ targets that achieve, at equilibrium, the intended targets of the

task. Identifying these specific λ targets crucially involves solving an (ill-posed) inverse problem. This is addressed in GEPPETO by a planning module that solves this inverse problem through an optimization procedure involving internal static forward models. Further details about this process are provided in Chapter 4.

The control architecture Since the particular shift of λ control variables is specified by the planning module prior to execution, the high level controller of GEPPETO corresponds to a purely feedforward control module. However, contrary to DIVA, this feedforward controller does not drive centrally specified movements of the effector, but controls the equilibrium parameters of the low level feedback loop of the effector system. The whole control architecture of GEPPETO can thus be seen as a high level feedforward module adjusting parameters of the low level feedback loop of the effector.

In this architecture, sensory feedback is only taken into account with respect to somatosensory information (muscle lengths) at the low level of the muscle stretch reflex. Although auditory feedback is known to play a role at the high control level, GEPPETO does not explicitly include this component for the moment. Auditory information is only taken into account as an internal prediction in the planning module of the controller. This implicitly assumes that auditory feedback is involved at the high control level during the learning and update of the internal model involved in the planner. GEPPETO implements a learning procedure of the relation between motor control variables (λ variables) and static acoustic characteristics of the speech signal that is based on a neural network approach relying on Radial Basis Functions. However, in its current state it does not include an update mechanism likely to relearn these relations in case of a perturbation altering it.

3.2 Variability, motor goals and perceptuo-motor interactions in speech motor control models

After reviewing the general context and the main approaches in speech motor control research, we will focus in this section on the specific questions that we address in this thesis: the variability of speech production, the sensory nature of speech motor goals, and the perceptuo-motor interaction. Our aim is to review how these questions have been addressed in the most acknowledged speech motor control models.

3.2.1 Variability of speech production

Variability is ubiquitous and has been well documented as a fundamental feature of speech production, which includes speaker-dependent and speaker-specific aspects (J. S. Perkell, 2013; J. S. Perkell & Klatt, 2014). In this thesis, we focus on speaker-specific aspects, which in turn include contextual and intrinsic components. Contextual variability of a speech sound might be associated either with the surrounding sounds (the so-called "coarticulation" phenomenon) or with prosodic factors such as speaking rate or stress. With intrinsic variability we mean the token-to-token variability that is observed across repetitions of a sound in the same phonetic and prosodic contexts. Contextual variability has been extensively studied experimentally and is a mandatory part of all the speech production models, since it is directly linked to linguistic factors and is then at the core of the interface between abstract phonology and physical phonetics. Intrinsic variability has been demonstrated by several experimental studies, but has been somewhat neglected in the speech production models, as it is considered to be a kind of noise, which has no significance to understand how speech production is organized in relation

to the linguistic levels. Yet it is an indisputable experimental fact, and it largely contributes to the naturalness of synthetic speech signal.

Contextual variability Contextual variability is usually associated with three basic properties of the speech production system and its interaction with speech perception: (1) redundancy; (2) categorical perception; (3) impact of articulatory mechanical inertia on speech movements due to the short durations of speech movements. Redundancy enables achieving with different motor commands the same characteristics in the space that characterize the motor task. Categorical speech perception is associated with the fact that variable acoustical patterns, and then vocal tract configurations, are associated with a unique phonological category. Put together these two properties enable the Central Nervous System selecting by purpose different motor commands, different articulatory configurations and different vocal tract shapes to produce the same motor goal, allophones of the same phoneme. This corresponds to motor equivalence in speech production (Perrier & Fuchs, 2015).

Motor equivalence has been extensively documented experimentally in speech production, and is evidenced for example by the fact that humans can easily speak with a pen maintained between the upper and the lower teeth. Motor equivalence is used in speech production in various situations such as adapting the articulation of a sound to the upcoming sounds (anticipatory coarticulation), producing a focus on a specific sound, or maintaining speech production accuracy under complex speaking conditions (for example speaking when walking, running or eating).

The short durations of speech movements make that speech articulators are essentially continuously moving, and rarely reach steady-state configurations, with the consequence that articulators' mechanical inertia impact the articulatory positions that will be used for the production of the up-coming sound. Mechanical inertia is a passive intrinsic property of the articulators. The Central Nervous System can take mechanical inertia into account and learn how deal with it, but it cannot control it or even less remove its influence. This results in noticeable variability within a given sequence of sounds when speaking rate increases, or for a given sound depending on the preceding sounds (the so-called "carry-over coarticulation").

In this context, speech production models can be differentiated in their ways to account for contextual variability according to the chosen implementation of redundancy, perceptually allowed variability, and mechanical inertia of the articulators.

In the Task dynamics model, as recalled above, speech motor goals are defined as dynamical attractors in an abstract task space which coordinates are the intended locations and degrees of the main constrictions in the vocal tract. An attractor in the task space is associated to each abstract phoneme. The movement toward a phoneme-related attractor is called a "gesture". This approach in the specification of the motor goals associated to a phoneme allows *de facto* at the abstract level of the intended vocal tract configuration a certain amount variability compatible with the correct perception of the phoneme. Indeed the attractor does not specify the full shape of the tongue, but only quite large anatomical regions of the vocal tract (lips, tongue tip, tongue dorsum, tongue rear) where constriction should occur. Redundancy is accounted for by the use of a geometrical articulatory model (the "plant"), which enables achieving a given constriction (i.e. a given position in the task space) with various articulatory positions (Saltzman, 1986, 1991). For instance, in a bilabial stop consonant as /p/, the abstract goal in the task space is a closed constriction at the lips, which does not specify anything for the tongue postures and enables different jaw openings, counterbalanced with different relative height of the upper and lower lips. The relation between the articulatory

positions and the location in the task space is implemented in the controller by "coordinative structures". Motor equivalence mechanisms emerge from the combination of the specification of the dynamical attractor in the task space and of the coordinative structures. For a given sequence of phonemes, anticipatory and carry-over coarticulation result from the specification over time of the dynamical attractors associated with the phonemes (the so-called "intergestural" coordination included in the gestural score) and the interarticulatory coordination implemented in the coordinative structures. Since the articulatory model is purely geometrical, the Task Dynamics model does not model articulatory inertia. However, the dynamical properties of the attractors in the task space is transmitted via the coordinative structures to the articulators, which enables the model simulating movement patterns that are similar to those that could be due to mechanical inertia, such as for example the reduction of a phoneme or a syllable when speaking rate increases (Browman & Goldstein, 1985).

In the DIVA model, basic speech units are phonemes and syllables. Motor goals are defined for each speech unit as large (time-varying for syllables) regions in the acoustic space and in the somatosensory space. The allowed variability compatible with the correction perception of the speech units is implemented via the size of these regions in the acoustic and somatosensory spaces. As in the Task Dynamics model, the redundancy is accounted for by the use of a geometrical articulatory model which enables achieving a same acoustic and somatosensory configuration with various articulatory positions. In the controller the relation between articulatory positions and acoustic and somatosensory variables is implemented in the context of a feedforward control scheme, using an inverse internal model that essentially describes how local articulatory changes are related to desired changes in the output variables. The combination of the definition of the motor goals and of the forward internal model implements motor equivalence. In particular the inverse internal model enables to impose no constraints on articulators which local variation has no impact on the output variables. For instance, in the case of the bilabial stop consonant /p/, for which lip closure, is crucial for the acoustic and somatosensory variables while tongue position has essentially no impact , large tongue variation is allowed and the combined dimension of jaw and lips controlling lip closure is crucially narrow. The DIVA model does not provide any account for mechanical articulatory inertia and does not enable any realistic account of speech variability such as phoneme reduction associated with variations in speaking rate.

In the DIVA model coarticulation is modelled in different ways. Within a syllable it would simply emerge from the inverse internal model: the articulatory pattern reached for a given phoneme will depend from the articulatory configuration reached for the preceding phoneme (carry-over coarticulation), as a combination of the size of the motor goals and the articulatory-to-acoustic and articulatory-to somatosensory sensitivity implemented in the inverse internal model. Thus, for instance the reached position for consonant /k/ would be different depending on whether the preceding phoneme is a front vowel as /i/ or a back vowel as /ɔ/ (Guenther, 1995). For frequently used syllables, anticipatory and carry-over coarticulation is implemented *de facto* by the specification of the time-varying motor goals (Guenther, Ghosh, & Tourville, 2006). In its original version (Guenther, 1995), DIVA accounted for anticipatory coarticulation with an approach in line with the anticipatory planning view. Guenther (1995) proposed a generalization of look-ahead models that accounts for anticipatory coarticulation (J. S. Perkell, 1980) "*by positing that movements for a feature of a later segment can start as long as the current segment and any intervening segments do not use that feature*"(Guenther, 1995). More recently in its extension to the GODIVA model anticipatory coarticulation has been implemented by serial activation/inhibition of the motor goals, as proposed in the serial-order model proposed

by (Lashley, 1951).

In the GEPPETO model, basic speech units are phonemes, but longer speech units are also considered at the level of the optimal planning (see below). Motor goals are defined as ellipsoid regions in the (F_1,F_2,F_3) formant space. All the configuration of these formant regions are assumed to be strictly equivalent from the perceptual point of view and to correspond to excellent realizations of the phonemes. As recalled above, speech movements within a speech sequence are generated in GEPPETO from the specification of the control variables at the successive targets of the sequence (corresponding to the successive phonemes) and by shifting the control variables from a target value to the next at a constant rate of shift. The model of the plant does not include a model of the jaw and of the lips. Hence the implemented redundancy is limited to the control variables shaping the tongue. Since the plant is a biomechanical model in which tongue shape is controlled by the activation of 6 independent muscles, redundancy is accounted for in the relation between the six muscular control variables (six λ parameters in the context of the Equilibrium Point Hypothesis of (Feldman, 1986) and the 3 formant values, which corresponds to 3 degrees of freedom. The indeterminacy associated with these degrees of freedom is solved at the level of a sequence of sounds (i.e. not for an isolated sound) on the basis of an optimal planning. This optimal planning minimizes in the space of the control variables the distance between the targets values, under the perceptual constraint that the target values correspond to excellent realizations of the intended phonemes of the sequence. This planning requires the learning of a static internal forward model associating motor control variables and (F_1,F_2,F_3) patterns. Thus, contextual variability results from this optimal planning.

Consequently, anticipatory coarticulation is modelled in GEPPETO by specifying the size of the sequence within which optimality applies. This approach enables GEPPETO to account for the existence of speech units larger than the phonemes (CV or CVC syllables, VCV sequences, ...), and for the influence of these larger units on coarticulation, thanks to the specification of the size of the sequence within which optimization applies (see Ma, Perrier, and Dang (2015); Perrier and Ma (2008))The proposed optimal planning does not take into account the serial order of the phonemes within the sequence (i.e. the /aki/ sequence is planned like the /ika/ or the /aik/ sequences). Hence, optimal planning also accounts for a part of the carry-over coarticulation. The remaining part of the carry-over coarticulation is generated during the execution of the motor task due to the intrinsic inertial properties of the biomechanical model of the tongue. Note that these intrinsic inertial properties are also responsible during the execution of the motor task for the variability associated with variation in speaking rate.

Intrinsic token-to-token variability

Theoretical origins The intrinsic component of variability is not specific to speech but is a fundamental aspect of all movements performed by biological systems: even the most skilled expert in any task would generally fail to repeat the exact same movement twice. While several studies have focused on this question from different descriptive (Lametti & Ostry, 2010; Shim, Latash, & Zatsiorsky, 2003) and theoretical perspectives (Churchland, Afshar, & Shenoy, 2006; Gordon, Ghilardi, Cooper, & Ghez, 1994), the origin of this fundamental component of movement variability remains controversial and poorly understood.

Theoretically, the origin of movement variability has been attributed to three main sources, resulting from the three main components involved in the control of movement: target specification, movement planning and movement execution (van Beers, 2007; van Beers, Haggard, & Wolpert, 2004). Furthermore, the variability arising from each of these sources may also have

different origins. For instance, in arm reaching movements, target specification corresponds to the localization of the target to be reached. Localization can be inaccurate, and hence variable, due to noisy sensory information or to deficits in information processing and estimation (which could be attributed to limitations in computational or memory resources, to the role of attention and fatigue, etc). Variability in movement planning can itself be attributed to noise in the neural networks in the brain (Faisal, Selen, & Wolpert, 2008), or to computational limitations leading to suboptimal planning (Beck, Ma, Pitkow, Latham, & Pouget, 2012; Wyart & Koechlin, 2016). Finally, movement execution can be corrupted by noise originating at different levels of the execution chain, from synapses or spikes in upper and lower motor-neurons, to the contractile response of muscles (Faisal et al., 2008).

The influence of these possible components of movement variability has been explored by a number of studies for arm and eye movements. However, as highlighted by van Beers (2007), these studies have proposed quite different conclusions. For instance, in the context of arm reaching movements, Gordon et al. (1994) and Churchland et al. (2006) essentially attributed variability to central movement planning, whereas van Beers et al. (2004) attributed it to movement execution. In the context of eye movements, Osborne, Lisberger, and Bialek (2005) and Osborne, Hohl, Bialek, and Lisberger (2007) attributed the main origin of variability in smooth pursuit to sensory information, and (van Beers, 2007) attributes variability of saccades to a combination of uncertainty in target localization and noise in movement planning and execution. Perhaps the best synthesis of the situation is provided by van Beers (2007), concluding that "*it is unlikely that a single source can explain all variability, although the relative contributions may vary across motor systems and movement types*" (van Beers, 2007, p. 8769).

Situation in speech Token-to-token variability has been observed and studied in a number of well-known experimental studies (Beckman et al., 1995; Folkins & Brown, 1987; Mooshammer, Perrier, Fuchs, Geng, & Pape, 2004; S. Perkell J. & Nelson, 1985). For an overview of the account of this kind of variability in speech motor control, we will adopt an evaluation based on the three sources of variability described above: at the level of localization, planning and execution.

The State Feedback Control applied to Task Dynamics follows an approach that has strong similarities with the hypothesis of variability in localization, since noise is explicitly assumed to corrupt the sensory signals that inform about the state of the system. However, noisy sensory signals are integrated with a Kalman filter approach that optimally estimates internal states of the speech production apparatus, discarding randomness. Hence, although noise is formally included in the model at the planning level, variability is discarded by the implementation of the optimal state estimator.

Another approach that has some similarities with the hypothesis of variability at the level of localization is the definition of large sensory regions for the specification of speech motor goals, as suggested by DIVA and GEPPETO. In this view, variability is attributed to the fact that the ultimate goals of speech production are abstract linguistic items that can be achieved by a large number of physical realizations. As a consequence, the intrinsic component of speech variability can be interpreted as an inaccuracy in target localization, yet not due to corrupted sensory signals, but to the fundamental abundant representation of speech motor goals. However, both in DIVA and GEPPETO this account of variability is subsequently removed at the planning level, since in both approaches the resulting abundance of possible solutions is resolved with a control strategy that selects always the same solution in a given context. Hence somehow paradoxically, both DIVA and GEPPETO include variability at the level of target specification, but preclude its intrinsic component due to the basic principles of their control strategies.

To our knowledge, only one very recent study has investigated intrinsic variability in speech production from the perspective of variation in planning (Tilsen, 2017). The focus of the study was the variability of the phase coupling between the consonantal onset and the vocalic nucleus within a CV syllable, in the theoretical context of the c-center theory (Browman & Goldstein, 1988). Experimental data were recorded in a session during which subjects had to repeat a syllable 400 times and received regular feedback about the consistency of their repetitions, in order to increase their exertion to the task. Variation in exertion was assumed to modulate the variability of the phase coupling between the syllable and experimental observations essentially confirmed this hypothesis. Using the model of coupled oscillators for the planning of syllable structure proposed by Nam, Goldstein, and Saltzman (2009), and providing random perturbations to the phase coupling between the oscillators, as a function of the assumed level of exertion, Tilsen (2017) was able to nicely reproduce some major aspects of the experimental findings. Even if it can be argued that the methodology used to study variability (interarticulatory phasing) includes *per se* the hypotheses that were tested at the level of the planning (planning based on coupled oscillators), this study provides an exciting example of how cognitive levels can influence the consistency of planning across repetitions in speech.

Intriguingly, no model of speech motor control has implemented or explored the role of execution noise in movement variability. This contrasts in particular with influential developments in computational motor control, in particular Stochastic Optimal Feedback Control (Todorov & Jordan, 2002), where optimization of movements in presence of execution noise has been suggested to account for a large number of behavioral features of movements. However, while certainly appealing, the implementation of these approaches in speech motor control raises issues about the online processing of feedback information due to the fact that the durations of the feedback delay and the speech gestures are similar. In this case only internal feedback predicted by forward models could be used and the efficacy of this approach is strongly dependent on the accuracy of this forward model. The question of this accuracy is particularly acute for speech production given the strong non-linearities between motor commands, articulatory configurations and acoustic patterns.

In summary, even though models of speech motor control explicitly acknowledge and include the intrinsic component of speech variability in their approaches, it remains formulated at a conceptual level and the concrete outcome of variability is often precluded by their computational approaches. The origin of token-to-token variability in speech production certainly has several components, at the level of target specification, planning and execution. However, the relative contribution of each of these components remains largely unknown.

In this thesis, in the context of the Bayesian reformulation of the planning part of the GEPPETO model, we will assess the potential influence of planning on token-to-token variability.

Multisensory motor goals The second question that we consider in this thesis is the sensory nature of speech motor goals. The sensory correlates of speech production are both auditory and somatosensory. Since sounds are a consequence of speech gestures, these two sensory modalities are redundant in unperturbed conditions. This raises questions about their functional involvement in the planning and monitoring of speech production: is only one useful, and if so, which one? Are they instead both useful, and if so, are they equivalent or complementary? The longstanding debate between auditory and articulatory theories of speech perception (Blumstein & Stevens, 1979; Fowler, 1991, 1996; Liberman & Mattingly, 1989; Schwartz, Basirat, Ménard, & Sato, 2012; Stevens, 1972, 1989) has led to an important number of experimental studies aiming at evaluating the involvement of auditory and somatosensory information in the control of speech (Brunner, Hoole, & Perrier, 2011; Fowler & Turvey, 1980; Gay, Lindblom, & Lubker,

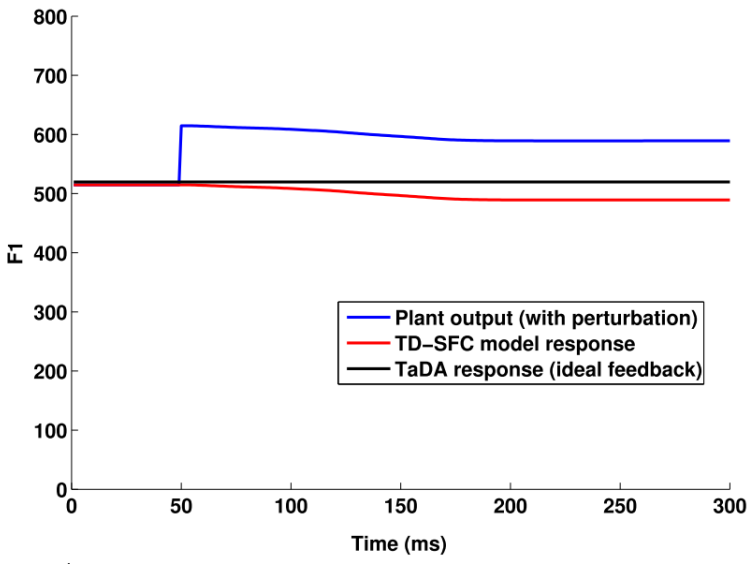

Figure 3.2: Compensatory behavior of the Task Dynamics – State Feedback Control model, from Ramanarayanan et al. (2016). *It can be seen that the model compensates the auditory perturbation and that compensation is incomplete.*

1981; Lametti, Nasir, & Ostry, 2012; Lubker, 1979; Savariaux, Perrier, & Orliaguet, 1995; Tremblay, Shiller, & Ostry, 2003). The paradigms of these studies consist in perturbing auditory or somatosensory feedback during speech production and assess whether subjects modify their control patterns in order to compensate for the perturbations. If compensation is observed it indicates that the perturbed sensory modality is involved in the control of speech.

These studies have provided evidence indicating that both auditory and somatosensory information appear to play a role in the control of speech, even speech production is maintained is postlingually deafened people (J. S. Perkell et al., 2000). However it remains unclear how these sensory information are actually involved. On the one hand, the DIVA model considers that speech motor goals are defined both in auditory and articulatory terms and thus both sensory information are used in order to monitor movements towards these two goals. On the other hand, the recent development of task dynamics, improved with state feedback control, is also able to include both auditory and somatosensory information. However, motor goals in Task Dynamics remain purely articulatory; and somatosensory and auditory information are only used in order to estimate the current state of the plant in the task of the space, i.e., the space of the constriction locations and sizes.

An important characteristics supporting the existence of both auditory and somatosensory goals is the fact that compensation to sensory perturbations in speech is never complete. This has been interpreted in the context of the DIVA model as resulting from the fact that both sensory targets cannot be reached in the presence of the perturbation (Villacorta, Perkell, & Guenther, 2007). Incomplete compensation would then result as a compromise between the two conflicting targets. This interpretation is certainly appealing, but it does not rule out the possible explanations suggested by the Task Dynamics model. Indeed, state feedback control is also able to account for incomplete compensation. In the case of perturbations, and prior to adaptation, the estimate of the plant integrates both sensory feedbacks and internal predictions that no longer agree. Depending on the relative weight attributed to sensory feedback and internal predictions, the estimate of the plant would be somewhere in between what would be expected from internal predictions alone and what would result from sensory feedback. Fig 3.2

illustrates the compensatory pattern obtained in Ramanarayanan et al. (2016), the blue curve presents the acoustic output of the model, with a 100 Hz perturbation applied at 50 ms. Around 50 ms later the model begins compensating for the auditory perturbation. Even if the authors did not discuss this effect, it can be seen that compensation reaches a plateau at around 170 ms that does not completely cancel out the perturbation (compensation reaches around 30% of the perturbation). Note however that after adaptation, and in the case of state estimates based only on auditory feedback, compensation would be complete again if the internal predictions are updated in order to agree with the perturbed auditory feedback. However, if both auditory and somatosensory feedbacks are taken into account in state estimation, incomplete compensation after adaptation would be recovered due to the mismatch between auditory and somatosensory feedback (Parrell et al., 2018).

Deepening in the involvement of auditory and somatosensory information in speech motor control, Lametti et al. (2012) provided evidence suggesting the existence of sensory preferences in speech production. In their study, both auditory and somatosensory perturbations were applied simultaneously. Subjects appeared to compensate for at least one of them, and a negative correlation was observed between amounts of compensation to each sensory perturbation. While in principle both the multisensory goal regions approach and the multisensory state estimation of state feedback control would be able to account for such sensory preferences, neither DIVA nor Task Dynamics are currently able to address this question in a way that is fully satisfactory for us. Indeed, somatosensory perturbations are performed by applying force perturbations that cannot be implemented on the geometrical descriptions of the plant considered in DIVA and Task Dynamics.

Perceptuo motor interactions The last question that we address in this thesis is the relation between perceptual and motor processes in speech. On the one hand, the fact that perception has an influence on motor processes is known and is well illustrated by compensations and adaptations to sensory-motor perturbations: perturbing perception has a clear influence on production. In this context, the DIVA model has already suggested a direct relation between perception and motor processes by relating the amount of compensation to perceptual acuity (J. S. Perkell, Lane, Ghosh, & Matthies, 2008; Villacorta et al., 2007). The state Feedback Control development of Task Dynamics would also account for such relation between sensory acuity and amount of compensation to auditory perturbations.

The idea that motor processes would play a role in speech perception has been introduced in particular by proponents of the motor theory of speech perception (Liberman & Mattingly, 1985), which suggests that speech perception would recover motor intentions from the acoustic sound. It has been suggested that the concept of gestures provided by the Task Dynamics and its link with phonological units proposed by the Articulatory Phonology offers a fruitful conceptual framework to understand how it could actually work: gestures could be the recovered motor intentions (Galantucci, Fowler, & Turvey, 2006). Recent experimental studies both in arm motor control research (Haith, Jackson, Miall, & Vijayakumar, 2009; Ostry, Darainy, Mattar, Wong, & Gribble, 2010) and speech (Lametti, Rochet-Capellan, Neufeld, Shiller, & Ostry, 2014; Nasir & Ostry, 2009; Schuerman, Nagarajan, McQueen, & Houde, 2017; Shiller, Sato, Gracco, & Baum, 2009) have highlighted a more intriguing phenomenon. Indeed, in the context of speech, shifts in perceptual boundaries have been shown to result from motor learning induced by perturbations of the auditory feedback (Lametti et al., 2014; Schuerman et al., 2017; Shiller et al., 2009), as well as perturbations of the articulatory gestures (Nasir & Ostry, 2009). To our knowledge, no current model of speech motor control is currently able to account for such phenomena since models are specifically designed for production but not for perception.

The Bayesian modeling framework developed in this thesis is particularly pertinent in this context since it enables to formulate perception and production as inferences processes performed from a shared set of knowledge. A consequence of this is that perception and production become naturally related in the framework. The last goal of this thesis is to apply this property for the interpretation of the experimental observations.

4 Conclusion

We have highlighted two main approaches in motor control. Normative approaches consider that regularities in behavior originate from properties of the task and that physical or biological constraints have limited impact on them. Both Task Dynamics and DIVA can be identified with this view, since in both models movements are essentially specified in the space corresponding to the task; biomechanical properties of the plant being assumed to have little influence on them. This view explains why these two modeling approaches have mainly focused on geometrical descriptions of the plant, since further levels of description are considered as secondary implementational details. The GEPPETO model crucially departs from this view, since its control strategy critically rests on fundamental physiological and biomechanical properties of the plant: its physical dynamics and the muscle stretch reflex. In this sense, the GEPPETO model is the only model of speech motor control that follows a more physical approach.

Beyond the differences that we have listed above, it is interesting to note that the three considered models have been developed in the aim to address different issues. This fact made that their respective contributions to the understanding of speech motor control are in slightly different domains. The DIVA model initially focused on trajectory formation and articulatory patterns, but currently mainly focus on the neural aspects of its control architecture. The Task Dynamics model has focused more on the importance of interacting dynamical systems in planning, variability and timing of speech production. The GEPPETO model has focused more on demonstrating the importance of biomechanics in the understanding of speech patterns.

We have presented an overview of the three main questions addressed in this thesis – variability, the sensory nature of speech motor goals, and perceptuo-motor interactions in speech – and have discussed the extent to which models of speech motor control have been able to account for them, and the different approaches that they have proposed. The contributions of this thesis with respect to these three questions are the following.

1. Concerning variability, our goal is to resolve the paradox that we observed in DIVA and GEPPETO: these models account for token-to-token variability at the level of target specification, but the principles underlying the selection of these targets preclude that this variability operates across repetitions. To do so, we reformulate the motor planning stage of the GEPPETO model in a Bayesian modeling framework that enables to formally characterize the abundance of possible realizations of speech items as uncertainty in motor planning. This approach does not discard an involvement of execution noise in speech production, but consider that its impact is not major compared to the variability that would originate at the levels of target specification and motor planning in speech.

2. Concerning the sensory nature of speech motor goals, our aim is to extend the GEPPETO model in order to take into account both auditory and somatosensory signals at the level of motor planning. To do so, we assume that both sensory signals are involved in motor planning with respect to two corresponding sensory characterizations of speech motor goals. This approach maintains the fundamental hypothesis of GEPPETO that speech

motor goals would be primarily auditory by considering that somatosensory motor goals are learned as somatosensory correlates of productions performed with respect to auditory goals.

3. Finally, a particularity of the Bayesian approach presented in this thesis enables us to address the question of perceptuo-motor interactions in an original way. Our Bayesian models formalize probabilistic representations of how knowledge would be structured and manipulated in the brain. In particular, we do not model production processes directly, but we identify them with the outcome of Bayesian inference based on the knowledge represented in the model. The same model thus also enables us to study perception processes from the same set of knowledge. In this context, perceptuo-motor interactions result as a natural consequence of the framework, since if production and perception processes rest on shared knowledge, they become naturally related in the framework.

Chapter 4

Overview of the GEPPETO model

1 Introduction

The key features of the GEPPETO model are nicely summarized by the meaning of its acronym: "**GE**stures shaped by the **P**hysics and by a **PE**rceptually oriented **T**argets **O**ptimization". More precisely, this statement summarizes three essential features of GEPPETO: (1) the fact that the physics of the speech apparatus are assumed to significantly influence the characteristics of the performed articulatory gestures; in order to account for this physical influence GEPPETO, in line with famous precursors (Kakita, Fujimura, & Honda, 1985; Kiritani, 1976; J. S. Perkell, 1974; Wilhelms-Tricarico, 1995), integrates a biomechanical description of the vocal tract; (2) the suggestion that speech motor goals are not continuous in time but discrete targets and, in line with Lindblom (1990), are essentially related to the perception of the sounds by listeners; (3) the idea that the control problem (resulting from the excess of degrees of freedom) is solved by optimization principles (Nelson, 1983).

These three key features also provide an outline for the presentation of GEPPETO. Since the physics of the speech apparatus are fundamental in GEPPETO, we begin by describing the biomechanical model on which GEPPETO rests (Section 2). This will allows us to introduce the main quantities manipulated in GEPPETO. In Section 3, we then present how GEPPETO specifies the speech task as concrete motor goals formulated in terms of these physical quantities that are related to perception. Finally, in Section 4, we present how GEPPETO solves the control problem in order to drive articulatory movements that achieve the intended goals of the speech task.

2 Specification of the plant

In order to take into account the physics of the speech apparatus, GEPPETO integrates a biomechanical description of the vocal tract. For simplicity, in this work we focus only on tongue movements, and consider a 2-dimensional biomechanical model of the tongue consisting of a finite element structure, representing the projection of the tongue on the mid-sagittal plane (Payan & Perrier, 1997; Perrier, Payan, Zandipour, & Perkell, 2003). Other more complex and more comprehensive biomechanical models of the speech production system have been developed in the last decade (Buchaillard, Perrier, & Payan, 2009; Dang & Honda, 2004; Gérard, Wilhelms-Tricarico, Perrier, & Payan, 2003; Hannam, Stavness, Lloyd, & Fels, 2008; Nazari, Perrier, Chabanas, & Payan, 2010; Pelteret & Reddy, 2012). We did not use them because they involve heavy computational loads, that would have not been compatible with the numerous simulations that have been necessary for the purpose of the present work. A number of studies have shown that, in spite of its simplicity, the 2-dimensional model is able to account for important kinematic characteristics of speech gestures (Payan & Perrier, 1997; Perrier & Fuchs, 2008; Perrier et al., 2003), see Section 5 for further discussion.

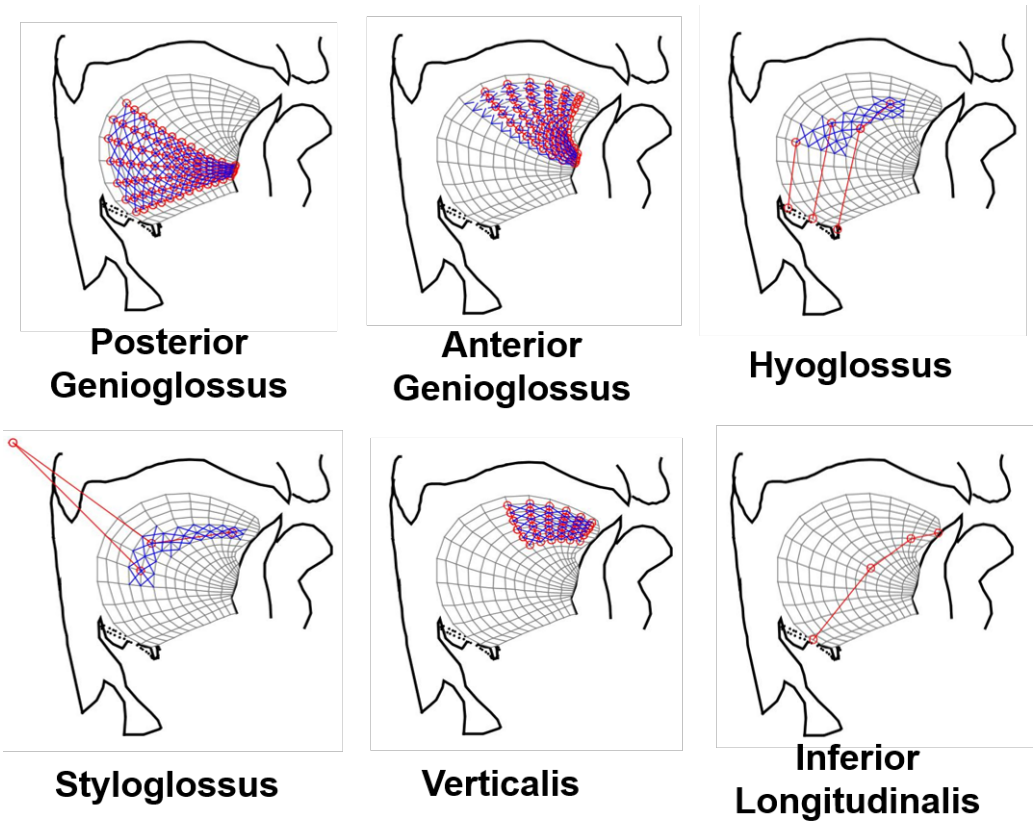

Figure 4.1: Biomechanical model of the tongue. *Colored lines correspond to fibers of each muscle. Crossed elements are the muscles elements and their elastic properties change with muscle activation.*

2.1 A 2-dimensional finite element structure for the tongue

The biomechanical model that we consider represents a sagittal view of the tongue as a 2-dimensional finite element structure. This finite element structure is a mesh that simulates the mechanical properties, inertia and Young modulus (equivalent to stiffness), of the soft tissues composing the tongue. Attached to this mesh, muscle fibers are included as actuators for shaping and driving movements of the tongue. Six principal muscles are considered in this version of the model. Fig 4.1 illustrates the mesh representing the tongue along with the fibers corresponding to each of the muscles considered in the biomechanical model.

Constraints acting on the tongue Movements and configurations of the mesh result from two kinds of constraints acting on the tongue; (1) internal forces, which can be further distinguished in terms of active forces due to muscle activation and passive forces due to the Young modulus characterizing the stress/strain relation, which increases in a muscle element as a function of its activation; (2) external forces, which may correspond to force contact with teeth or the palate, gravity, external perturbations or inertial forces (as during running). In the current version of the model, only contact is implemented.

Generation of muscle force: the Equilibrium point Hypothesis Muscle forces are generated in the model according to the Equilibrium Point Hypothesis: the activation of each

muscle results from the interaction between a centrally specified threshold activation parameter λ, corresponding to the muscle length above which active muscle force is generated (Feldman, 1986), and the length and length change rate of the muscle. More precisely, the force $f_k(t)$ generated by muscle k at time t is specified as:

$$f_k(\lambda_k,t) \;\coloneqq\; \rho_k \left[\exp\left(c_k\ \alpha_k(\lambda_k,t)\right) - 1\right], \tag{4.1}$$

where c_k is a form parameter accounting for the gain of the feedback from the muscle to the motor neurons pool and ρ_k a magnitude parameter directly related to force-generating capability (related to the muscle cross-sectional area). $\alpha_k(\lambda_k,t)$ is the muscle activation at time t corresponding to:

$$\alpha_k(\lambda_k,t) \;\coloneqq\; \left[l_k(t) - \lambda_k(t) + \gamma_k \dot{l_k}(t)\right]^+, \tag{4.2}$$

where $\lambda_k(t)$ is the threshold activation parameter of muscle k at time t, $l_k(t)$ is the actual length of muscle k, $\dot{l_k}(t)$ the muscle lengthening velocity and γ_k a damping coefficient providing stability to the system (Payan & Perrier, 1997). $[x]^+$ is defined as

$$[x]^+ \;=\; \coloneqq \begin{cases} x & \text{if } x \geqslant 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

and implements the requirement that muscle activation is either positive or zero.

Eqs (4.1) and (4.2) state that the force generated by each muscle depends both on the state of the muscle (muscle length and velocity) and on the value of its λ parameter. If muscle length l_k is smaller than or equal to λ_k, no active muscle force is generated. Muscle force can be generated in two ways: (1) by stretching the muscle, such that muscle length becomes longer than the activation threshold λ; (2) by specifying an activation threshold λ that is shorter than muscle length. The first alternative corresponds to muscle force induced by the stretch reflex in response to a muscle elongation generated by an external influence. The second alternative enables active control of muscle force (yet not direct control since reflex mechanisms are also involved).

2.2 Acoustic outputs

The acoustic output of GEPPETO is generated in two main steps, from the 2-dimensional shape of the vocal tract determined by the upper contour of the tongue model and the external contours of the vocal tract, i.e. the hard palate, the soft-palate including the velum, and the posterior wall of the pharynx, which are all supposed to be fixed and rigid.

In the first step an estimation of the area-function of the vocal tract is computed. The area-function is the variation of the cross-sectional area of the vocal tract, from the glottis to the teeth. Its estimation corresponds to an extension of the description of the vocal tract, from the 2D representation in the mid-sagittal plane to a volumic representation. This is done thanks to a model (Perrier, Boë, & Sock, 1992) that accounts for the relation between the sagittal dimension of the vocal tract and its cross-sectional area. This model implements non-linear functions, which have been inferred on the basis of 3D data collected from the vocal tract of a French speaker producing the main French vowels in a CT-scan, and from the cast of the vocal tract of a male cadaver. Former studies in the lab (Wu, Badin, Cheng, & Guerin, 1987) have shown that a description of the area-function as a series of short tubes, which lengths do not exceed 5 mm, is accurate enough to enable reliable acoustic signals in the frequency-range below 6 kHz. In addition appropriate natural time variation of the speech signal can be

synthesized if the area-function is up-dated every 10ms. Hence for a sequence of movements of the biomechanical model of the tongue, the output of this first step is a sequence of area-functions made of 44 tubes of equal length smaller than 5 mm, and sampled at 100 Hz.

In the second step, the time-varying spectral characteristics of the sound produced by the movements of the biomechanical model are computed thanks to a harmonic analog of the vocal tract that has been designed by Badin and Fant (1984). This model computes for each area-function (every 10ms then) the corresponding 4 formant values associated with the tongue position. Thus, after this two steps, starting from a sequence of control variables generating tongue movements, we obtain a sequence of 4 formants in the range [0 6] kHz sampled at 100 Hz.

If a synthetic speech signal is needed, the acoustic model provided by Story (Story, Laukkanen, & Titze, 2000; Story & Titze, 1995) is used. This model transforms each area-function into a frame of acoustic signal using Kelly-Lochbaum modelling (Kelly, 1973). The successive frames are then concatenated with usual signal processing techniques in order to generate an acoustic signal for the whole duration of the tongue movement.

2.3 Driving tongue movements: the control variable

The control variable in GEPPETO corresponds to the 6-dimensional vector of λ parameters:

$$m(t) \;\coloneqq\; (\lambda_1(t), \ldots, \lambda_6(t)). \tag{4.4}$$

Each configuration of m leads to a particular equilibrium configuration of the tongue, in which the stresses generated by all the active and passive forces counterbalance each other. Changing the values of m modifies the forces generated by each muscle, such that the configuration of the tongue is no longer in equilibrium and starts moving towards a new equilibrium configuration in which forces balance each other again. Therefore, tongue movements correspond to transitions between equilibrium configurations driven by changes in λ control variables. Furthermore, the particular trajectory and velocity profile of the tongue during its transition from one equilibrium configuration to the other results from the interaction between the evolution of λ control variables and the particular configuration of the tongue. This is particularly relevant for the production of different speech rates as described in Section 4.

3 Specification of the task

GEPPETO considers that speech motor goals are not continuous in time but are characterized as a sequence of discrete targets units. Yet, speech can be performed at various speech rates and with different prosodic and lexical stress patterns. Therefore, defining a speech task in GEPPETO corresponds to specifying "what" should be produced and "how". The "what" question aims at specifying the identity of the intended target units, and most crucially, what characterize them in concrete perceptual terms. The "how" question aims at specifying speech rates and stress patterns and how they are characterized in concrete motor terms.

3.1 "What": Nature of speech motor goals

3.1.1 Speech as a sequence of phonemes

GEPPETO considers that speech can be specified as a sequence of fundamental phonological units $(\phi^1, \cdots, \phi^n)$, corresponding to phonemes of the considered language (Dell, 1986). In the

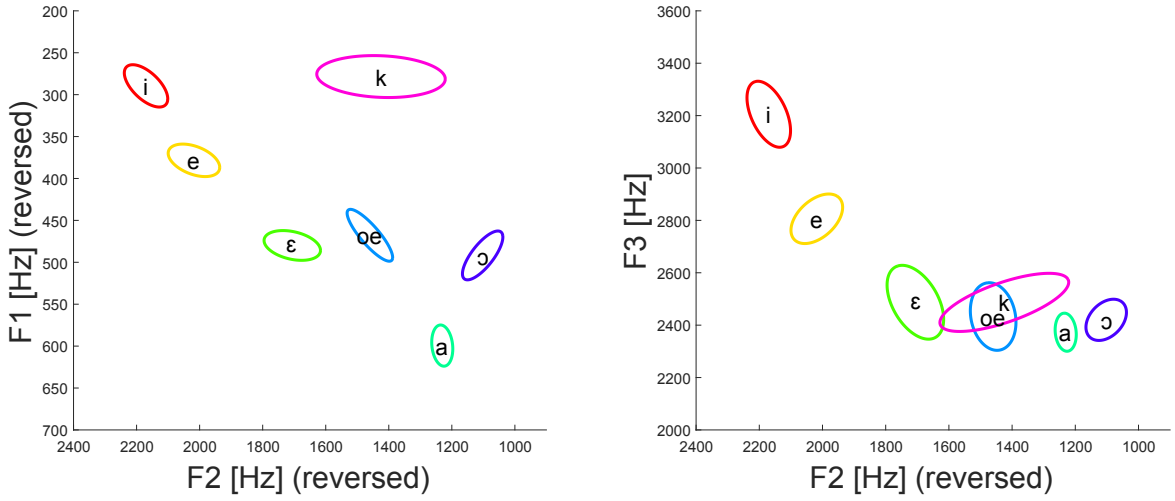

Figure 4.2: Projections of the 3-dimensional dispersion ellipsoids corresponding to each target region characterizing phonemes. Left: (F_2,F_1) plane; Right: (F_2,F_3) plane.

present case, given the limitations of the biomechanical model describing only the tongue, we focus on French phonemes that do not involve jaw or lip movements:

$$\phi^i \in \{/\text{i}/, /\text{e}/, /\varepsilon/, /\text{a}/, /\text{oe}/, /\text{ɔ}/, /\text{k}/\}. \tag{4.5}$$

3.1.2 Phonemes specified by auditory target regions

GEPPETO considers that phonemes are essentially related to the perception of the sounds by listeners and therefore are characterized and controlled in a way that is strongly related to the perceptually relevant features of these sounds. In the case of vowels, these features have been identified with the three first peaks of the spectral envelope of the acoustic signal, and therefore the perceptually relevant space considered by GEPPETO is the 3-dimensional formant space (F_1, F_2, F_3).

Furthermore, GEPPETO characterizes phonemes as target regions in this 3-dimensional auditory space. The characterization of speech motor goals in terms of regions, rather than point-like targets, aim at accounting for the natural range of variability observed in speech production, an assumption in the same line of DIVA (Guenther et al., 1998)

Phoneme target regions in GEPPETO are identified with dispersion ellipsoids – characterized by their centers μ_A^ϕ and covariance matrices Γ_A^ϕ for each phoneme ϕ – established on the basis of phoneme production experiments (Calliope, 1984; Ménard, 2002; Robert-Ribes, 1995) and adapted to the acoustic maximal vowel space of the model (Perrier, Ma, & Payan, 2005; Winkler, Ma, & Perrier, 2011). Fig 4.2 represents the projection of these regions on the (F_2,F_1) and (F_2,F_3) planes.

3.2 "How": Speech rate, stress and clarity

Speech can be performed at different speech rates, with different levels of clarity and with different patterns of stress. These can correspond to segmental features (lexical stress) or suprasegmental features (prosodic stress) (Beckman et al., 1995). This useful versatility is enabled in GEPPETO by the fact that the same sequence of phonemic goals can be performed with different timings and with different levels of muscle coactivation (different configurations of

λ control variables can lead to the same tongue shape, and therefore the same acoustic output, but with different levels of generated forces).

3.2.1 Specification of speech rate

GEPPETO implements speech rate by specifying the intended duration of movements for each element i in the sequence as $(T^1, \cdots, T^n)$. However, it is possible to perform the same sequence with the same speech rate but holding longer phonemes and moving faster between phonemes, or moving slowly between phonemes and holding phonemes for shorter times. Therefore, each duration T^i is further specified in terms of transition τ^i, and holding h^i phases (see Fig 4.4). While currently there is no principle for the specification of transitions and holding times in GEPPETO, transitions are rather associated with speech rate while holding times are associated with emphasis of phonemes.

3.2.2 Specification of effort levels

Controlling the level of force is a requirement to control accuracy. It is also useful to account for stress patterns, either prosodic stress such as focus or lexical stress (Perrier, Lœvenbruck, & Payan, 1996). GEPPETO characterizes every articulatory configuration at the targets with a corresponding level of effort w. These levels of effort are associated with the levels of total active muscle force ν, defined as the sum of active forces generated by each muscle at its equilibrium configuration:

$$\nu(m) \quad = \quad \sum_{i=1}^{6} f_i(\lambda_i) \ . \tag{4.6}$$

GEPPETO defines levels of effort w by categorizing total forces ν into three levels Winkler et al. (2011):

$$w \quad \in \quad \{\text{“}\textit{Weak}\text{”}, \text{“}\textit{Medium}\text{”}, \text{“}\textit{Strong}\text{”}\}. \tag{4.7}$$

However, since muscle force capacity is highly muscle dependent (Brand, Pedersen, & Friederich, 1986; Pruim, De Jongh, & Ten Bosch, 1980) and since phonemes involve different patterns of recruited muscles (Baer, Alfonso, & Honda, 1988; Buchaillard et al., 2009; K. Honda, 1996; Waltl & Hoole, 2008), levels of effort are not simply defined with respect to absolute amplitude of muscle force, but also with respect to the identity of the intended phoneme. This is a way to take into account physiological characteristics of muscles in the definition of effort, and distinguish phonemes, such as /i/, that requires the activation of intrinsically strong muscles, such as the Genioglossus or the Styloglossus, as compared to phonemes, such as /a/ that is associated with a weaker muscle, the Hyoglossus. Fig 4.3 represents, for each phoneme, the corresponding ranges of total muscle force ν assigned to each level of effort w.

3.2.3 Dealing with accuracy: balance between speaking rate and effort level

Speech rate and effort levels are two independent features in the model. However, in order to achieve accuracy at high speech rates, effort levels need also to be taken into account. Indeed, recall that the basic principle of movement generation in the model rests on the fact that λ changes induce a difference between the equilibrium tongue shape (as controlled by the λ parameters) and the actual tongue shape (Eqs (4.1) and (4.2)). This results in the generation of a force that will tend to move the tongue towards its new equilibrium. The faster the change

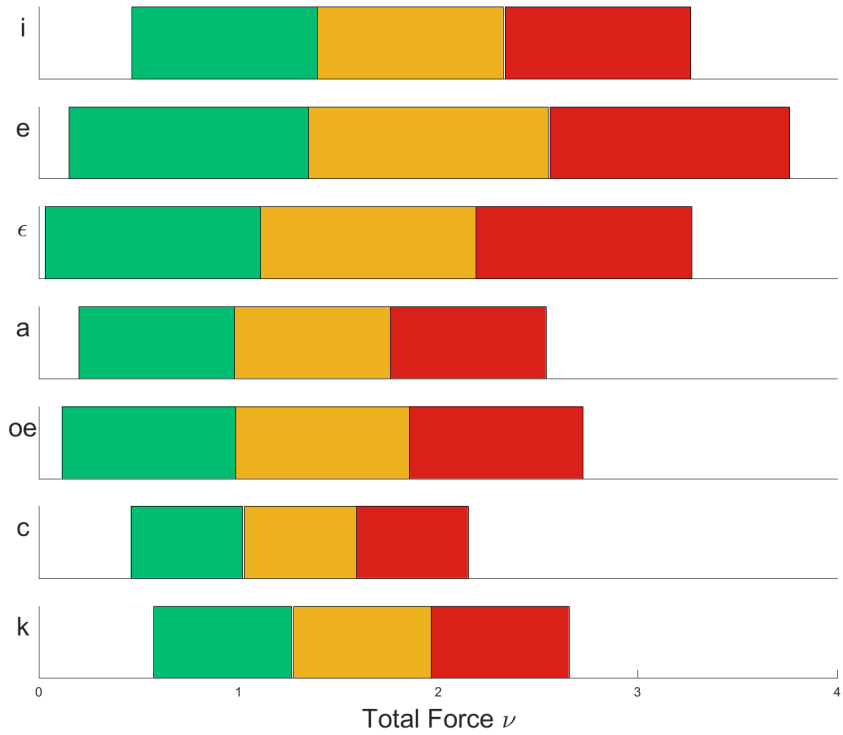

Figure 4.3: Ranges of total muscle force ν assigned to each level of effort w for each phoneme. "Weak" levels in green, "Medium" levels in yellow, "Strong" levels in red.

in λs, the greater the distance between the equilibrium configuration and the actual tongue shape. The reaction force towards the equilibrium configuration increases as a consequence of this greater distance, but this increase could be insufficient, when the distance is too large, for the tongue to reach the final equilibrium configuration within the time constrained by speaking rate. As a consequence, it is useful in this situation to jointly specify effort level and speaking rate: a high level of accuracy at a high speaking rate requires the generation of high level of forces. Note however that this joint specification of effort level along with intended speaking rate would correspond to a planning module that is not currently implemented in the model.

3.2.4 Summary

In summary, the task in GEPPETO is defined as a sequence of n triplets:

$$\{(\phi^1, w^1, T^1), \cdots, (\phi^n, w^n, T^n)\} \ . \tag{4.8}$$

These triplets specify the identity of each of the intended phonemes along with the intended levels of effort w^i and speech rate formulated as duration times T^i with which the sequence is intended to be produced. Phonemes ϕ^i are defined as auditory goal regions in formant space (F_1, F_2, F_3), and effort levels w^i are defined as phoneme-dependent ranges of total generated forces ν. Duration times are further composed of transition and holding phases $T^i = (\tau^i, h^i)$.

4 Specification of the controller

4.1 The control problem

The aim of the controller in GEPPETO is to specify the evolution of the control variable, $m(t)$, such that the resulting movements of the tongue satisfy the requirements of the task, i.e., to produce acoustic signals that reach the intended sequence of auditory target regions with the intended levels of effort and with the intended speech rate. However, the speech task can be achieved by many possible control patterns. First, since both phoneme categories and effort levels are characterized as regions rather than specific points, the speech task can be achieved by many different tongue configurations at targets. Second, the transition between two tongue configurations at targets can be performed in many possible ways. Specifying the evolution of the control variable is therefore an ill-posed problem and a strategy must be provided in order to address this redundancy problem.

4.2 Principle of the control strategy

To begin with, note that a particularity of GEPPETO, compared with other models of speech motor control, is that the solution to the control problem does not involve inverse kinematics (Jordan & Rumelhart, 1992) or inverse dynamics (Kawato, 1999). The fundamental reason is that GEPPETO does not specify desired trajectories or action plans in distal, auditory or articulatory, terms. Instead, GEPPETO directly solves the control problem at the proximal level of the control variable, and trajectories in articulatory and auditory spaces result from the interaction between the evolution of the control variable and the biomechanical properties of the tongue.

So, how does GEPPETO solve the control problem at the level of the control variable? GEPPETO's strategy is structured in two steps: (1) use optimization principles in the space of the control variable in order to select a sequence of control targets $(m^{*1}, \cdots, m^{*n})$ for which the corresponding equilibrium configurations of the tongue satisfy the intended phonemes and effort levels of the task; (2) specify the evolution of the control variable $m(t)$ such that it reaches the selected control targets $(m^{*1}, \cdots, m^{*n})$ at the intended timings $(T^1, \cdots, T^n)$ specified by the speaking rate of the task.

4.2.1 First step-planning: optimal selection of a sequence of control targets

In the first step of its strategy, GEPPETO addresses the first problem described earlier: many different configuration of control targets $(m^{*1}, \cdots, m^{*n})$ can produce the intended phonemes and effort levels of the task.

Optimality principle. In order to address this redundancy problem, GEPPETO assumes that the selection of control targets $(m^{*1}, \cdots, m^{*n})$ is solved by an optimality principle that intends to minimize displacements in control space.

Constraints. In order to ensure that the optimal selection of control targets $(m^{*1}, \cdots, m^{*n})$ satisfies the requirements of the task, optimization must be performed under two constraints. The first constraint is perceptive, and intends to ensure that the resulting auditory consequences achieve the intended auditory target region. The second is an effort constraint that intends to ensure that generated forces ν agree with the specification of effort levels.

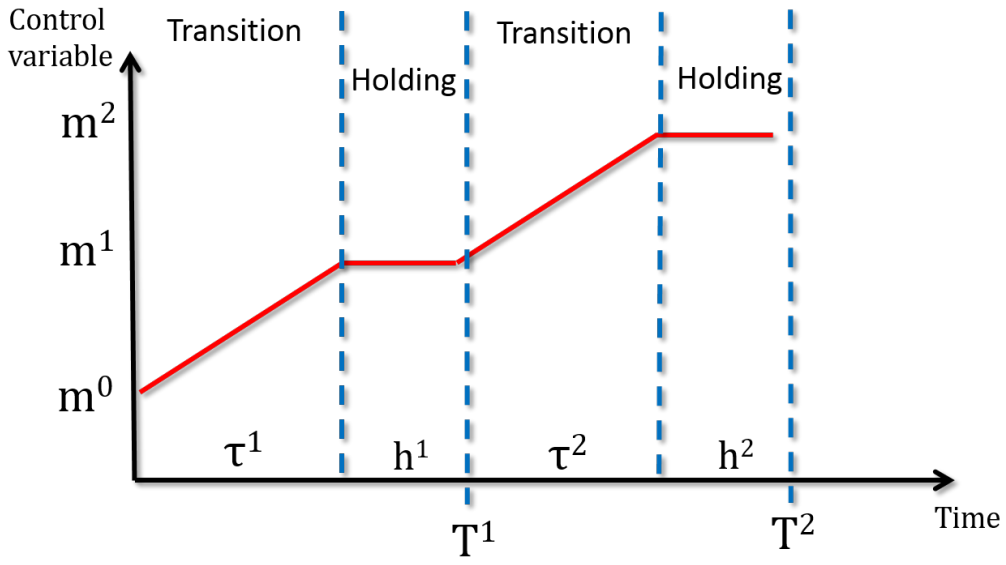

Figure 4.4: Evolution of the control variable $m(t)$ as a sequence of linear transitions between control targets (m^{*1},m^{*2}). m^0 *corresponds to* λ *values in the rest configuration of the tongue, from which movements begin.*

Internal "static" forward models. The optimality principle is defined in the proximal space of the control variable, however constraints are defined with respect to the distal requirements of the task. In order to relate the proximal control space to the two distal requirements of the task, GEPPETO assumes the existence of two internal forward models, ρ_ν and ρ_a, that respectively predict the resulting level of force ν and auditory consequences (F_1,F_2,F_3) of control variable m. These internal models are considered to be "static" as they associate values of the control variable m with their corresponding consequences at equilibrium configurations.

4.2.2 Second step-execution: specify the time evolution of the control variable

In this second step, GEPPETO addresses the second problem described earlier: the selected control targets $(m^{*1}, \cdots, m^{*n})$ can be attained by many possible time evolutions of control variable $m(t)$, even when the intended timings $(T^1, \cdots, T^n)$ are specified. GEPPETO's solution to this problem rests on a principle of simplicity: the evolution of the control variable $m(t)$ is considered to be as simple as possible and is therefore defined as a sequence of linear transitions between the selected control targets $(m^{*1}, \cdots, m^{*n})$, with transition and holding times $(\tau^1 + h^1, \cdots, \tau^n + h^n)$ specified by the intended speaking rate.

Note that the actual tongue trajectory, its shape and its timing in terms of transitions and holding phase, results from the interaction between the specified time evolution and the biomechanical characteristics of the tongue. See Fig 4.4 for an illustration of the resulting evolution of the control variable $m(t)$ for a sequence of two targets.

4.3 Concrete implementation of the control strategy

In the previous section we have defined the general principle of GEPPETO's control strategy. Now we describe its concrete implementation in computational terms.

4.3.1 Planning of control targets: Optimization and constraints

The goal of this first step is to find the optimal sequence of control targets $(m^{*1}, \cdots, m^{*n})$ that minimize displacements in control space, under the constraint that the resulting tongue configurations lead to forces and auditory consequences that satisfy the intended effort and auditory target regions.

GEPPETO implements this optimization with a gradient descent algorithm based on a cost function $\mathcal{F}_C$ including one term, C_M, implementing the optimality criterion and two additional terms, C_A and C_ν, implementing the perceptual and effort constraints respectively:

$$\mathcal{F}_C(m^1, \cdots, m^n) \;\coloneqq\; \mathcal{C}_M(m^1, \cdots, m^n) + \sum_{i=1}^{n} \left[\mathcal{C}_A^{\phi^i}(m^i) + \mathcal{C}_\nu^{w^i \phi^i}(m^i) \right] \;. \tag{4.9}$$

The solution to the control problem is thus specified as:

$$(m^{*1}, \cdots, m^{*n}) \;=\; \min_{(m^1, \cdots, m^n)} \mathcal{F}_C(m^1, \cdots, m^n) \;. \tag{4.10}$$

The algorithm is initialized by selecting a random set of control targets that satisfy both the perceptual and effort constraints.

Optimality criterion. For a three-phoneme sequence, the optimality criterion is defined by the perimeter of the triangle defined by the three considered control targets:

$$\mathcal{C}_M(m^1, m^2, m^3) \;\coloneqq\; ||m^2 - m^1|| + ||m^3 - m^2|| + ||m^3 - m^1||. \tag{4.11}$$

For a general n-phoneme sequence the proposed cost function would correspond to the perimeter of the corresponding $(n-1)$-simplex defined by the n control targets in the 6-dimensional control space. For the present 3-phoneme case, the 2-simplex corresponds to the triangle mentioned above. Rigorously, influence of every phoneme of the sequence on every other one would be rather modeled by a cost function involving distances between every pair of phonemes. In order to avoid the corresponding quadratic combinatorial growth of the number of terms in the cost function, its definition has been simplified into the one presented here.

Perceptual and effort constraints. The aim of the perceptual and effort constraints is to ensure that selected control targets satisfy the requirements of the task. More specifically, the spectral properties (F_1, F_2, F_3) of the resulting acoustic signals must lie within the ellipsoids regions assigned to each phoneme ϕ in the sequence, and the generated forces ν must be within the ranges of the intended effort level w for each phoneme in the sequence. This is implemented by including two costs that vanish whenever the predicted acoustic signal belong to the intended region and goes to infinity (in practice a large number) otherwise:

$$\mathcal{C}_A^{\phi}(m) \;\rightarrow\; \begin{cases} 0 & \text{if } \rho_a(m) \;\in\; \mathcal{E}(\mu_A^{\phi}, \Gamma_A^{\phi}) \\ \infty & \text{otherwise} \end{cases} \tag{4.12}$$

$$\mathcal{C}_\nu^{w\phi}(m) \;\rightarrow\; \begin{cases} 0 & \text{if } |\rho_\nu(m) - \mu_\nu^{w\phi}| \leqslant \sigma_\nu^{w\phi} \\ \infty & \text{otherwise.} \end{cases} \tag{4.13}$$

Concerning the auditory cost in Eq (4.12), $\mathcal{E}(\mu_A^{\phi}, \Gamma_A^{\phi})$ is the set of points inside the ellipsoid region in auditory space corresponding to phoneme ϕ; μ_A^{ϕ} is the 3-dimensional vector specifying its center and Γ_A^{ϕ} its corresponding 3×3 covariance matrix; $\rho_a(m)$ is the predicted acoustic

consequence resulting from the control variable m. Concerning the effort cost in Eq (4.13), $\mu_\nu^{w\phi}$ and $\sigma_\nu^{w\phi}$ are respectively the center and width of the interval of forces corresponding to the effort level w for phoneme ϕ; $\rho_\nu(m)$ is the predicted force associated to the control variable m. GEPPETO assumes that auditory and force predictions are performed by two internal models, ρ_a and ρ_ν, that are described below.

Internal models The perceptual and effort constraints rely on the ability to associate values of control variables to their predicted forces and auditory consequences. These predictions are implemented by two internal models ρ_ν and ρ_a, that result from learning processes that generalize, from a limited number of examples, the relations between values of the control variable m and corresponding forces ν, in the case of ρ_ν, and corresponding formants, in the case of ρ_a. These models are considered to be "static" as they associates values of the control variable m and corresponding forces or auditory consequences at equilibrium configurations. GEPPETO implements internal models through Radial Basis Function (RBF) (Poggio & Girosi, 1989) learned through classical supervised learning.

5 Experimental results accounted by GEPPETO

In spite of the simplified description of tongue biomechanics included in the biomechanical tongue model (2D description and linear mechanics), GEPPETO was able to reproduce interesting characteristics of speech movements, based on a target-to-target control, without any inversion mechanisms aiming at tuning the commands in order to reproduce this characteristics. Payan and Perrier (1997) have thus shown that the model was able to generate velocity profiles similar to those observed on human speakers. Perrier et al. (2003) have shown that the model naturally generates the complex looping articulatory trajectories observed in different languages in vowel-velar consonant-vowel sequences. Brunner et al. (2011) have also shown that the model was able to reproduce the variability of these looping patterns observed in Korean velar stops depending on their voicing status, by changing the articulation location of the velar stops as human speakers do. Perrier and Fuchs (2008) have also shown that in the tongue trajectories generated with GEPPETO the relation between tangential velocity and the curvature corresponds to a power law just as the real tongue movements measured for speakers of different languages (see also Tasko and Westbury (2004)). Such a power law corresponds to a characteristics observed in many human movements (Viviani & Terzuolo, 1982), and sometimes attributed to some jerk minimization (Viviani & Flash, 1995).

Part I

Intrinsic variability

Chapter 5

Bayesian reformulation of GEPPETO

1 Introduction

The starting point of this thesis aims at addressing the computational dilemma raised by variability in speech production. The dilemma is most clearly illustrated in DIVA and GEPPETO, for which target regions rather than target points are assumed in order to account for speech variability. By defining target regions, both models explicitly include variability at the representational level of the task. However, they rely on computational strategies for which the resulting abundance of solutions is an ill-defined and degenerated situation that must be resolved. The result is that, while the abundance of target regions enable to account for several aspects of contextual variability (as motor equivalence, coarticulation and speed rate effects), their computational approaches are unable to formally address the intrinsic component of speech variability, since once the context is fixed, abundance must be discarded in order to deduce a unique and invariant solution to their control problem.

In this chapter we intend to address this dilemma by formulating speech motor planning in an alternative computational framework in which the abundant characterization of speech motor goals does not lead to ill-defined control problems. We address speech motor planning as an inference problem formulated in a probabilistic framework. This enables to formally account for variability as uncertainty at the motor planning level of the speech production task. We illustrate this approach by reformulating the planning module of the GEPPETO model in a Bayesian modeling framework.

Our modeling approach is based on the Bayesian Programming framework (Bessière, Laugier, & Siegwart, 2008; Bessière, Mazer, Ahuactzin, & Mekhnacha, 2013; Lebeltel, Bessière, Diard, & Mazer, 2004), which is a systematic and generic methodology for the definition of Bayesian programs. Applied to cognitive modeling at Marr's algorithmic level (Marr, 1982), this becomes Bayesian algorithmic modeling (Diard, 2015). Bayesian programming and Bayesian algorithmic modeling allow the definition of arbitrarily complex models, thanks to modularity and hierarchical constructs. These models encode knowledge in the form of sets of probability distributions that constitute joint probability distributions; applying Bayesian inference then yields expressions for computing terms of interest, which we call "questions", not contained in the model as elementary pieces of knowledge. These questions model process that we want to simulate.

We begin this chapter by introducing in more details the Bayesian Programming methodology, in Section 2, by first briefly reviewing the rules of probability on which Bayesian modeling rests, and then presenting the Bayesian Programming methodology with a simple example. We then present our Bayesian reformulation of GEPPETO step-by-step. We first focus on simulating motor planning for the production of single phonemes, in Section 3. We then build up on this first model and extend it, in Section 4, in order to address motor planning for the production of phoneme sequences.

2 The Bayesian Programming framework

2.1 Language and rules of probability theory

Variables and probabilities Probability theory, in its subjectivist interpretation, is an extension of logic. The relation between the subjectivist and frequentist interpretations of probabilities is a large topic in itself, which we will not discuss here (Jaynes, 2003). However, this view of probabilities as an extension of logic has a few practical implications, among which are a number of particularities in notation and terminology. For instance, instead of "random variables" we will refer to "probabilistic variables", instead of noting sets of variables as X_1, X_2 we will refer to their conjunction $X_1 \wedge X_2$, or for simplicity $X_1 X_2$, etc.

While in logic reasoning is based on propositions that are either false or true (in binary terms: 0 or 1), in probability theory reasoning is performed with continuous degrees of belief about the truth of logical propositions (represented with real numbers from 0 to 1). For instance, if you are almost certain that a coin you just flipped landed on heads, $P([Coin = head]) = 0.99$.

It is then convenient to group together possible outcomes of events in sets called "probabilistic variables", provided these sets are mutually exclusive (no two events can be true simultaneously) and exhaustive (at least one of them is true). Reprising our coin flipping example, the assertion would be about the state of the coin after landing, which can be either head or tail, and would be represented by a binary probabilistic variable: $C = \{head; tail\}$. Your confidence about the truth of your assertion is represented by a real number between 0 and 1; 0 being certainly false, and 1 being certainly true. If coin flipping is performed with a fair coin, you have maximal uncertainty before the coin lands and therefore you would assign the same probability, to both possibles outcomes: $P([C = head]) = 0.5$ and $P([C = tail]) = 0.5$. $P(C)$ is therefore a probability distribution over states of C, representing your knowledge about each of these possibles outcomes.

Conditional probabilities Belief is in general relative to some previous state of knowledge. For instance, in the previous example you had maximum uncertainty because you were told that the coin was fair. If you knew that the coin was unfair, for instance biased towards tails, your uncertainty would have been reduced and you would have attributed more probability to tails than heads. Let us represent the possibles states concerning the bias of the coin by an additional variable B, corresponding to the possibles amounts of bias of the coin towards tails. B is a continuous variable in [0, 1] where 0 stands for a coin that always falls on heads, 1 for a coin that always falls on tail, 0.5 for a fair coin and all other possible values for the different proportions of possible bias. The fact that your belief about the outcome of a throw depends on your knowledge about the bias of the coin is formulated as a conditional probability distribution $P(C \mid B)$. In the present case, your knowledge of the relation between the bias of the coin and the outcome of the throw would be formulated as:

$$P([C = tails] \mid [B = b]) \quad = \quad b. \tag{5.1}$$

Independence and conditional independence Variables A and B are said to be independent when $P(A \mid B) = P(A)$, for all values of A and B. Furthermore, variables A and B are said to be independent, conditionally on C when $P(A \mid B\ C) = P(A \mid C)$, for all values of A, B and C.

Rules of probability theory Up to now we have illustrated how knowledge and uncertainty are translated into probabilistic variables and probability distributions. Now we introduce what are the basic rules that enable us to manipulate this knowledge.

The sum or normalization rule is the first fundamental rule. It states that the sum of probabilities of all possible outcomes of a given probabilistic variable is 1. In mathematical terms:

$$\sum_{c \in C} P([C = c] \mid B) \;=\; 1, \tag{5.2}$$

which, we write in short as:

$$\sum_{C} P(C \mid B) \;=\; 1. \tag{5.3}$$

In order to anticipate and avoid confusions, we draw attention to the notation that is employed here. Variables can be discrete or continuous. Usually, one writes P for probability distributions over discrete variables and p for probability densities over continuous variables. For simplicity, we choose not to make this distinction here. Similarly, all summations and integrals are denoted by the sign $\sum$, even when rigorously it is the $\int$ sign that should be used for continuous variables.

The product or conjunction rule enables to compute joint probability distributions over conjunction of variables. For two variables it corresponds to:

$$\begin{aligned} P(C\ B) &= P(B)P(C \mid B) && (5.4)\\ &= P(C)P(B \mid C), && (5.5) \end{aligned}$$

and its generalization to n variables:

$$P(X_1\ \ldots\ X_n) \;=\; P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2\ X_1)\ldots P(X_n \mid X_{n-1}\ \ldots\ X_1) \tag{5.6}$$

The marginalization rule can be derived from Eqs (5.4) and (5.2) as:

$$\sum_{C} P(C\ B) \;=\; P(B), \tag{5.7}$$

Bayes rule can be derived from Eqs (5.4) and (5.5), provided that $P(B) \neq 0$:

$$P(C \mid B) \;=\; \frac{P(C)P(B \mid C)}{P(B)}, \tag{5.8}$$

or equivalently:

$$P(C \mid B) = \frac{P(C\ B)}{P(B)} = \frac{P(C\ B)}{\sum_{C} P(C\ B)}, \tag{5.9}$$

where the denominator in the last term was obtained by applying the marginalization rule of Eq (5.7).

$$\text{Program} \left\{ \begin{array}{l} \text{Description} \left\{ \begin{array}{l} \text{Specification} \left\{ \begin{array}{l} \text{Variables} \\ \text{Decomposition} \\ \text{Parametric forms} \end{array} \right. \\ \text{Identification} \end{array} \right. \\ \text{Question} \end{array} \right.$$

Figure 5.1: Structure of a Bayesian Program.

Eq (5.9) is our most useful rule. It enables to compute conditional probabilities whenever the joint probability distribution over a set of probabilistic variables is known. Eq (5.9) further generalizes to $(n+1)$ variables, $\{A_1, A_2, \cdots, A_n, B\}$ as:

$$P(A_1 \mid B) = \frac{P(A_1\ B)}{P(B)} = \frac{\sum\limits_{/\{A_1,B\}} P(A_1\ \cdots A_n\ B)}{\sum\limits_{/B} P(A_1\ \cdots A_n\ B)}, \tag{5.10}$$

where $\sum\limits_{/X}$ denotes summing over all variables but those included in the set X.

We also consider conditional probabilities of more than two variables. In this case the product rule generalizes, which can also be obtained in a similar way:

$$P(A_1 \mid A_2\ B) = \frac{P(A_1\ A_2\ B)}{P(A_2\ B)} = \frac{\sum\limits_{/\{A_1,A_2,B\}} P(A_1 \cdots A_n\ B)}{\sum\limits_{/\{A_2,B\}} P(A_1 \cdots A_n\ B)}, \tag{5.11}$$

$$P(A_1 \mid A_2\ A_3\ B) = \frac{P(A_1\ A_2\ A_3\ B)}{P(A_2\ A_3\ B)} = \frac{\sum\limits_{/\{A_1,A_2,A_3,B\}} P(A_1 \cdots A_n\ B)}{\sum\limits_{/\{A_2,A_3,B\}} P(A_1 \cdots A_n\ B)}. \tag{5.12}$$

2.2 The Bayesian Programming framework with a simple example

We have briefly introduced the main concepts and rules of probability theory that underlie our modeling work. We now introduce the Bayesian Programming methodology (Bessière et al., 2008, 2013; Lebeltel et al., 2004) that provides us with a template for structuring the definition of Bayesian models. Fig 5.1 illustrates the structure of this template, which is organized in two main steps: (1) the "Description", corresponding to the precise definition of the model, and (2) the "Questions", corresponding to the inferences to be computed from the model definition. The "Description" gathers all the assumed knowledge and hypotheses concerning the relevant quantities – as well as their relations – involved in the problem. It corresponds to translating all the relevant prior knowledge into the language (and with the rules) of probabilities, in particular by quantifying incompleteness as uncertainty. The "Questions" correspond to the actual reasoning with the knowledge formulated in the first step. Results obtained at this step are directly derived as a consequence of applying the rules of probability theory to the joint probability distribution.

To present the methodology with a simple example, we begin by considering the perceptual constraint of GEPPETO. The aim of this constraint is to ensure that the produced sounds correctly achieve the intended auditory regions characterizing phonemes. The constraint is

implemented as a term in the cost function that penalizes sounds that lie outside the intended target region. Our aim is to reformulate this perceptual constraint in probabilistic terms by identifying it with the outcome of an inference process that enables to categorize sounds into phoneme identity, high probability values being attributed only to sounds that achieve the intended categorical target regions.

2.2.1 Description

The Bayesian Programming methodology begins with a "Description" step that is itself decomposed into "Specification" and "Identification" stages. The aim of the "Specification" is to define the joint probability distribution that gathers all the knowledge and hypotheses concerning quantities involved in the problem. This is achieved by first defining the probabilistic **variables** involved in the model, then proposing a **decomposition** of their joint probability distribution, and finally assigning **parametric forms** to all terms in this decomposition. All these knowledge are extracted from the fundamental hypotheses of GEPPETO, presented in Chapter 4 and summarized in Fig 5.3.

Variables To begin with, we first need to identify what are the relevant quantities involved in the problem and how they should be represented in probabilistic terms. In the present case we are interested in the categorization of sounds into phoneme identity. The relevant quantities in this categorization are therefore phonemes and auditory stimuli, which we represent by variables Φ_c and A_c respectively.

$\boldsymbol{A_c}$ represents auditory stimuli as specified by GEPPETO. Hypothesis H_2 in Fig 5.3 specifies speech sounds as points in the 3-dimensional formant space. Therefore, we define A_c as a continuous 3-dimensional vector variable, $A_c = (F_1, F_2, F_3)$. The domain of this variable is expressed in Hertz and is assumed to be bounded between extreme values attained by the simulations of the biomechanical model of the tongue: $\mathcal{D}_A = \mathcal{D}_{F_1} \times \mathcal{D}_{F_2} \times \mathcal{D}_{F_3}$, where $\mathcal{D}_{F_1} = [100, 700]$, $\mathcal{D}_{F_2} = [800, 2200]$ and $\mathcal{D}_{F_3} = [2300, 3500]$.

$\boldsymbol{\Phi_c}$ is a discrete variable representing the set of phoneme considered in GEPPETO (hypothesis $H_{4.1}$). An additional "no-phoneme" category (denoted by $\emptyset$) is further assumed in order to take into account all auditory configurations that do not fall within any of the phonemic categories defined in GEPPETO. The values taken by variable Φ_c are thus labeled by:

$$\Phi_c \in \{ /\text{i}/, /\text{e}/, /\varepsilon/, /\text{a}/, /\text{oe}/, /\text{ɔ}/, /\text{k}/, \emptyset \} .$$

Decomposition Having defined variables in the model, we now need to to specify their joint probability distribution $P(\Phi_c\ A_c)$. This is performed by decomposing it as a factor of simpler terms, using the product rule in Eq (5.4):

$$P(\Phi_c\ A_c) = P(\Phi_c)\ P(A_c \mid \Phi_c). \tag{5.13}$$

Parametric forms The decomposition in Eq (5.13) is further specified by defining the forms of the factors that it features.

$\boldsymbol{P(\Phi_c)}$ corresponds to the knowledge that we have *a priori* about phonemes Φ_c. The mathematically simplest case is to make no assumption about them, and thus to define $P(\Phi_c)$ as a uniform probability distribution:

$$P(\Phi_c) = \frac{1}{8}. \tag{5.14}$$

$\boldsymbol{P(A_c \mid \Phi_c)}$ corresponds to the knowledge about the expected sounds A_c associated with each phoneme Φ_c. We assume this knowledge to be represented by Gaussian probability distributions, truncated to the range of the domain $\mathcal{D}_A$:

$$P([A_c = a] \mid [\Phi_c = \phi]) = \begin{cases} \frac{1}{Z_\phi(\kappa_A)} \mathcal{G}(a\ ;\ \mu_A^\phi,\ \kappa_A^2 \Gamma_A^\phi) & \text{if } a \in \mathcal{D}_A \\ 0 & \text{otherwise} \end{cases} \quad (5.15)$$
$$\text{for } \phi \in \{\ /\text{i}/,\ /\text{e}/,\ /\varepsilon/,\ /\text{a}/,\ /\text{oe}/,\ /\text{ɔ}/,\ /\text{k}/\ \}\ ,$$

where κ_A is a parameter introduced in order to modulate the dispersion of the probability distributions (see left panels of Fig 5.2 for illustration), and where $Z^\phi(\kappa_A)$ is a normalization factor due to the truncation of the distributions to the domain $\mathcal{D}_A$.

However, recall the special case of the "no-phoneme" category $\emptyset$: to represent maximal uncertainty, the probability distribution corresponding to this category is chosen to be uniform over the domain $\mathcal{D}_A$. Denoting by $|\mathcal{D}_A|$ the volume of this domain, we define:

$$P([A_c = a] \mid [\Phi = \emptyset]) = \begin{cases} \frac{1}{|\mathcal{D}_A|} & \text{if } a \in \mathcal{D}_A \\ 0 & \text{otherwise.} \end{cases} \quad (5.16)$$

Identification The final stage in the "Description" corresponds to the identification of the particular values of parameters characterizing these forms, usually determined from experimental data. In the present case these parameters are the means μ_A^ϕ, covariance matrices Γ_A^ϕ and parameter κ_A characterizing each Gaussian distribution. Means μ_A^ϕ and covariance matrices Γ_A^ϕ can be directly identified to the centers and matrices defining the quadratic forms characterizing the ellipsoid target regions in GEPPETO (hypothesis $H_{4.1}$). We keep parameter κ_A purposely unspecified in order to implement particular assumptions in the model; its value will be given for each result.

2.2.2 Question

The "Description" completely defines the Bayesian model, i.e, it completely specifies the joint probability distribution $P(A_c\ \Phi_c)$. The next stage is to use the knowledge included in the joint probability distribution in order to address particular inference questions. These questions are identified with probability distributions of interest. In the present case, we are interested in the categorization of an auditory input a into phoneme identity Φ_c, which is formulated in probabilistic terms as the probability distribution $P(\Phi_c \mid [A_c = a])$. This inference is obtained from Bayes rule in Eq (5.9), together with the decomposition of the joint probability distribution in Eq (5.13):

$$\begin{aligned} P(\Phi_c \mid A_c) &= \frac{P(\Phi_c) P(A_c \mid \Phi_c)}{\sum_\phi P([\Phi_c = \phi]) P(A_c \mid [\Phi_c = \phi])} \\ &= \frac{P(A_c \mid \Phi_c)}{\sum_\phi P(A_c \mid [\Phi_c = \phi])}, \end{aligned} \quad (5.17)$$

where $P(\Phi_c)$ was simplified from the last line since we assume it to be uniform.

Eq (5.17) completely specifies the probability distribution $P(\Phi_c \mid A_c)$ identified to the categorization question. Right panels in Fig 5.2 illustrates the likelihood function $P(\Phi_c \mid A_c) = f(A_c)$ resulting from the inversion of auditory characterizations $P(A_c \mid [\Phi_c = k])$.

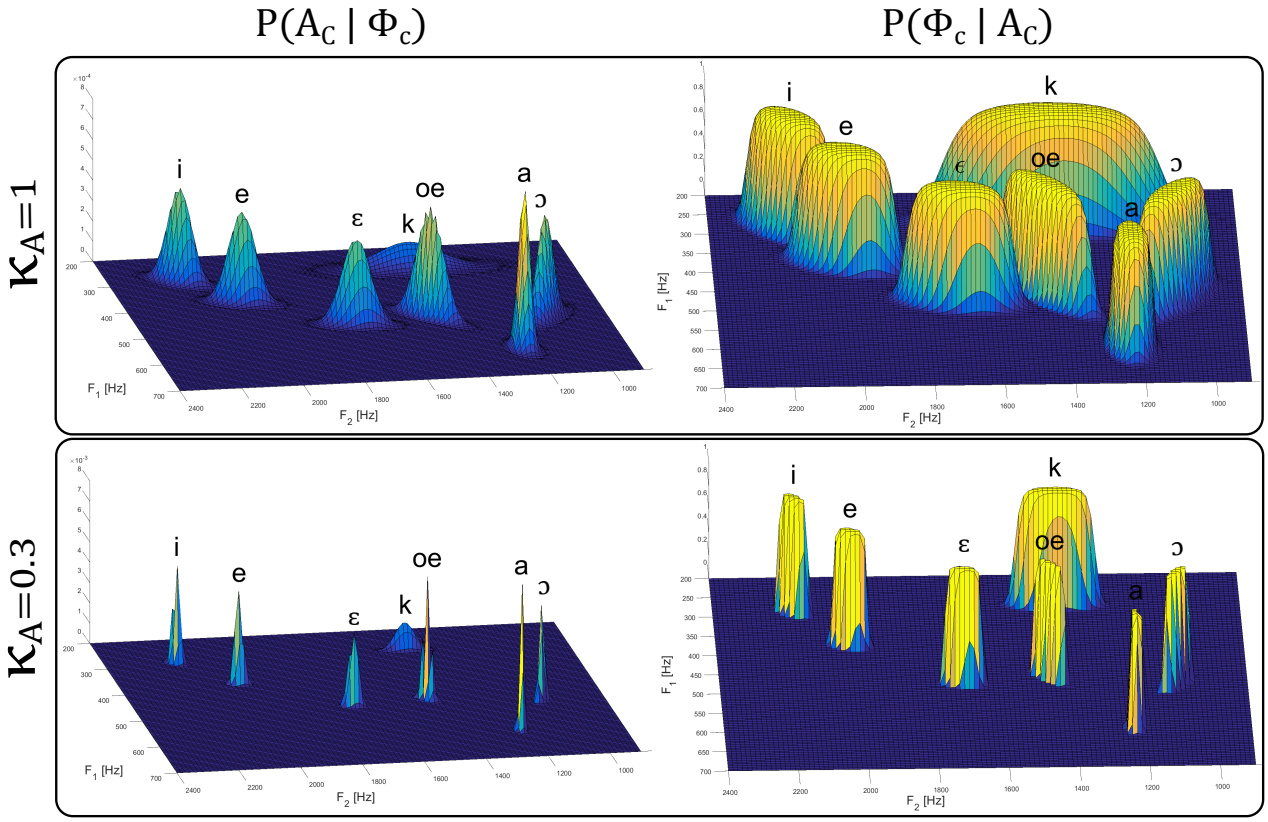

Figure 5.2: Auditory characterizations of phonemes $P(A_c \mid \Phi_c)$ (left panels) and corresponding categorization functions $P(\Phi_c \mid A_c)$ (right panels), in the (F_1,F_2) plane. Top panels: $\kappa_A = 1$; Bottom panels $\kappa_A = 0.3$.

3 Bayesian model for the production of single phonemes with intended levels of effort

In the previous section we introduced the Bayesian Programming methodology by illustrating it with the probabilistic reformulation of the perceptual constraint in GEPPETO. Now we go a step further in the reformulation of GEPPETO and formulate the inference of values of the control variable in GEPPETO for the production of single phonemes. The model presented in this section builds up on the previous introductory example by including motor knowledge and the effort constraint associated with generated forces. We begin by presenting the construction of the model in Section 3.1 and then, in Section 3.2, we formulate and compute the motor planning question, i.e., the inference of control variables for the production of a single phoneme. This Section is adapted from Patri, Diard, and Perrier (2015); Patri, Perrier, and Diard (2016).

3.1 Model description

3.1.1 Variables

Variables in this model are simply extracted from the key ingredients of GEPPETO described in Chapter 4 and summarized in Fig 5.3.

M represents the control variable in GEPPETO. According to hypothesis H_1 in Fig 5.3 the control variables corresponds to the six λ parameters specifying the threshold activation

The plant

(H_1) **The control variable** in GEPPETO is a 6-dimensional vector of λ parameters specifying the threshold activation length of each muscle: $m = (\lambda_1, \cdots, \lambda_6)$.

(H_2) **Acoustic signals** are represented in 3-dimensional formant space: $a = (F_1, F_2, F_3)$.

(H_3) **The total level of force** of each tongue configuration is defined as the sum of forces generated by each muscle in their equilibrium configuration.

The task

(H_4) **The speech task is defined as a sequence** $\{(\boldsymbol{\phi^1}, \boldsymbol{w^1}, \boldsymbol{T^1}), \cdots, (\boldsymbol{\phi^n}, \boldsymbol{w^n}, \boldsymbol{T^n})\}$, each triplet (ϕ^i, w^i, T^i) specifying intended phonemes ϕ^i, effort levels w^i and durations T^i (composed of transitions τ^i and holding times h^i).

($H_{4.1}$) **Phonemes** are characterized in auditory terms by particular target regions, in 3-dimensional formant space, identified with dispersion ellipsoids of parameters μ_A^ϕ and Γ_A^ϕ (Fig 4.2). Considered phonemes are: $\phi \in \{$/i/, /e/, /ɛ/, /a/, /oe/, /ɔ/, /k/$\}$.

($H_{4.2}$) **Effort levels** are associated with particular ranges of generated forces depending on the identity of the intended phoneme (Fig 4.3). Considered effort levels are: $w \in \{$"Weak", "Medium" , "Strong"$\}$.

The controller

(H_5) **The control strategy** is divided into a **planning stage** and an **executions stage**.

(H_6) **The aim of the planning stage** is to specify a discrete sequence of control targets $\{m^{*1}, \cdots, m^{*n}\}$ that generates acoustic consequences and total forces that agree, respectively, with the target regions characterizing the intended phonemes ($H_{4.1}$), and the range of forces characterizing the intended effort levels ($H_{4.2}$). The planning stage is performed by an optimization process performed by a gradient descent algorithm resting on the following features:

($H_{6.1}$) **A perceptual constraint** $\mathcal{C}_A$ ensures that the selection of control targets result in acoustic consequences (predicted via the internal model ρ_a ($H_{6.4}$)) that agree with the intended auditory target regions.

($H_{6.2}$) **An effort constraint** $\mathcal{C}_\nu$ ensures that the selection of control targets results in levels of force (predicted via the internal model ρ_ν ($H_{6.5}$)) that agree with the intended level of effort.

($H_{6.3}$) **A motor constraint** $\mathcal{C}_M$ aims at favoring minimal displacements in control space. It is implemented as a motor cost identified, in the case of a three phoneme sequence, with the perimeter of the triangle defined by the sequence of control variables $\{m^1, m^2, m^3\}$.

($H_{6.4}$) **An auditory-motor internal model** $\boldsymbol{\rho_a}$ represents the knowledge of the mapping from control variable $m = (\lambda_1, \cdots, \lambda_6)$ to the resulting acoustic consequences $a = (F_1, F_2, F_3)$.

($H_{6.5}$) **A motor-to-force internal model** $\boldsymbol{\rho_\nu}$ represents the knowledge of the mapping from control variable $m = (\lambda_1, \cdots, \lambda_6)$ to the resulting generated force ν.

(H_7) **The execution stage** performs the time evolution of the control variable $m(t)$ in the form of linear transitions between the specified control targets $\{m^{*1}, \cdots, m^{*n}\}$ (Fig 4.4), with transitions and holding times corresponding to the intended durations T^i specified by the task (H_4).

Figure 5.3: **Summary of fundamental hypotheses in GEPPETO**

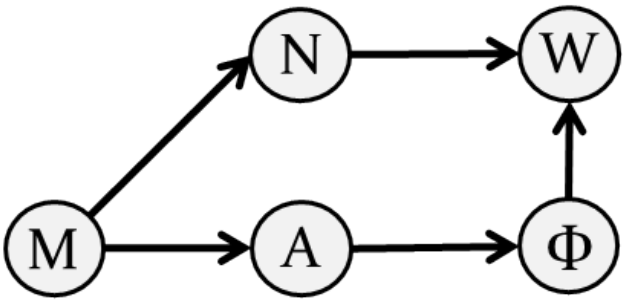

Figure 5.4: Graphical representation of the single phoneme model. *The diagram represents the decomposition of the joint probability distribution given by Eq (5.19). Nodes represent variables in the model and arrows represent their dependencies.*

length of each muscle. M is therefore a continuous 6-dimensional vector variable defined as $M = (\lambda_1, \ldots, \lambda_6)$. The domain of M corresponds to the range of values of each parameter λ_i for which the bio-mechanical model attains its equilibrium configurations within the ranges constrained by the vocal tract boundaries.

Φ and A are phonemic and auditory variables defined in the same way as Φ_c and A_c in the previous introductory model in Section 2.2. Their labels have been modified in order to avoid confusions. Φ is therefore a discrete variable with values

$$\Phi \;=\; \{\ /\text{i}/,\ /\text{e}/,\ /\varepsilon/,\ /\text{a}/,\ /\text{oe}/,\ /\text{ɔ}/,\ /\text{k}/,\ \emptyset\ \}\ .$$

and A is a continuous vector variable, $A = (F_1, F_2, F_3)$, in the same domain as defined Section 2.2.

N is a continuous scalar variable representing the total level of force associated to each tongue configurations (H_3).

W is a discrete variable representing the level of effort generated at the target articulation of a phoneme. Its elements are the 3 levels of effort considered in GEPPETO (hypothesis $H_{4.2}$): $W = \{$ "Weak", "Medium", "Strong"$\}$.

3.1.2 Decomposition

We now define the structure of the Bayesian model, by specifying the joint probability distribution over the five above variables. Following the product rule given in Eq (5.6), an exact decomposition of the joint probability distribution $P(M\ A\ \Phi\ N\ W)$ is given by:

$$\begin{aligned} &P(M\ A\ \Phi\ N\ W) \\ &\quad = \ P(M)\ P(A \mid M)\ P(\Phi \mid A\ M)\ P(N \mid \Phi\ A\ M)\ P(W \mid N\ \Phi\ A\ M). \end{aligned} \tag{5.18}$$

This exact decomposition can be further simplified by taking into account conditional independence assumptions resulting from GEPPETO. According to hypothesis $H_{4.1}$, phonemes are assumed to be fully characterized in auditory terms. Therefore, Φ can be assumed to be independent of M conditioned on the knowledge of A, such that $P(\Phi \mid A\ M) = P(\Phi \mid A)$. Next, the total force N generated by each tongue configuration in the biomechanical model is uniquely specified by the motor variable M. Therefore, we can simplify $P(N \mid \Phi\ A\ M)$ into $P(N \mid M)$.

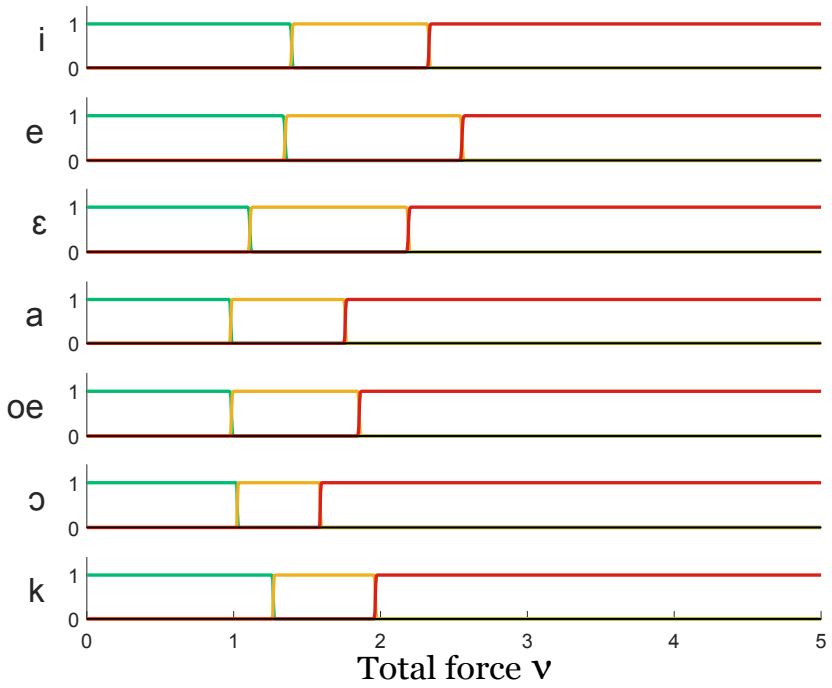

Figure 5.5: Likelihood functions corresponding to each $P(W \mid [N = \nu]\ \Phi)$ for each phoneme and each effort levels. *Green curves correspond to effort level W =“Weak”, yellow to W =“Medium” and red to W =“Strong”.*

Finally, according to hypothesis $H_{4.2}$, the level of effort W is only characterized by the knowledge of Φ and N, such that $P(W \mid N\ \Phi\ A\ M) = P(W \mid N\ \Phi)$. With these simplifications, Eq (5.18) becomes:

$$\begin{aligned} & P(M\ A\ \Phi\ N\ W) \\ & = \ P(M)\ P(A \mid M)\ P(\Phi \mid A)\ P(N \mid M)\ P(W \mid N\ \Phi). \end{aligned} \tag{5.19}$$

Figure 5.4 illustrates the Bayesian network representing this decomposition.

3.1.3 Parametric forms

$\boldsymbol{P(M)}$ is the prior probability distribution over motor control variables M. Since no prior knowledge is assumed about this variable, $P(M)$ is defined as a uniform probability distribution over domain $\mathcal{D}_M$:

$$P(M) := \begin{cases} \frac{1}{|\mathcal{D}_M|} & \text{if } M \in \mathcal{D}_M \\ 0 & \text{otherwise.} \end{cases} \tag{5.20}$$

$\boldsymbol{P(A \mid M)}$ **and** $\boldsymbol{P(N \mid M)}$ represent the knowledge of the motor-to-auditory and motor-to-force mappings relating control variables M to their auditory and force consequences, A and N, respectively. They correspond to the internal models ρ_A and ρ_ν implemented by Radial Basis Functions (RBF) networks in GEPPETO (hypotheses $H_{6.4}$ and $H_{6.5}$). Denoting by $\rho_A(m)$ and $\rho_\nu(m)$ the auditory and force outputs associated to motor input m, the corresponding probability distributions are assumed to be deterministic and are given by:

$$P([A = a] \mid [M = m]) \ := \ \delta(a - \rho_A(m)) \tag{5.21}$$

$$P([N = \nu] \mid [M = m]) \ := \ \delta(\nu - \rho_\nu(m)) \tag{5.22}$$

where δ denote Dirac distributions, translating the fact that $P([A = a] \mid [M = m])$ (resp. $P([N = \nu] \mid [M = m])$) is zero unless $a = \rho_A(m)$ (resp. $\nu = \rho_\nu(m)$).

$\boldsymbol{P(\Phi \mid A)}$ corresponds to the probability of assigning phoneme Φ to the given auditory signal A. It can be interpreted as the auditory categorization of phonemes, as provided by the model defined in Section 2.2. This categorization model translates in probabilistic terms the perceptual constraint associated with the auditory regions characterizing each phoneme in GEPPETO (hypothesis $H_{6.1}$). Therefore, we identify $P(\Phi \mid A)$ in this model with the outcome of the previous model $P(\Phi_c \mid A_c)$:

$$P(\Phi \mid A) \quad \coloneqq \quad P(\Phi_c \mid A_c). \tag{5.23}$$

Recall that the variance of each Gaussian distribution in $P(\Phi_c \mid A_c)$ is controlled by a parameter κ_A that allows to control the precision of the categorization task (see Fig 5.2, for illustration).

$\boldsymbol{P(W \mid N\ \Phi)}$ characterizes the relation between the level of effort W and the total force N generated for a given phoneme Φ. It can be seen, for each phoneme, as the categorization of the generated forces N into one of the three level of efforts W, and is thus identified with the effort constraint of GEPPETO (hypothesis $H_{6.2}$). As for $P(\Phi \mid A)$, this term is specified by a sub-model that performs this categorization based on the probability distributions $P(N \mid W\ \Phi)$ of forces N generated with respect to each level of effort W and for each phoneme Φ. $P(N \mid W\ \Phi)$ are assumed to be 1-dimensional Gaussian distributions, with parameters specified by the effort constraint implemented in GEPPETO. The resulting $P(W \mid N\ \Phi)$ for each effort W and each phoneme Φ, are illustrated in Fig 5.5.

3.2 Question: Inference of values of the control variables M

Having specified the joint probability distribution $P(M\ A\ \Phi\ N\ W)$, we now formulate the questions to be solved by the Bayesian model. As the problem is to infer motor control variables M producing a desired phoneme Φ with a given level of effort, the approach consists in computing the probability distribution over M, conditioned on the specified value of Φ and W. The corresponding probability distribution, $P(M \mid \Phi\ W)$, is obtained by applying Bayes rule in Eq (5.9):

$$\begin{aligned} P(M \mid \Phi\ W) &= \frac{\sum_{A,N} P(M\ A\ \Phi\ N\ W)}{\sum_{M,A,N} P(M\ A\ \Phi\ N\ W)} \\ &\propto \sum_{A,N} P(M\ A\ \Phi\ N\ W), \end{aligned} \tag{5.24}$$

where the proportionality symbol "$\propto$" on the last line accounts for the fact that the denominator does not depend on M, for a given value of Φ and W. Using the decomposition given by Eq (5.19) and including the constant term $P(M)$ into the proportionality symbol, we obtain:

$$\begin{aligned} &P([M = m] \mid \Phi\ W) \\ &\quad \propto \sum_{A,N} P(A \mid [M = m])\ P(\Phi \mid A)\ P(N \mid [M = m])\ P(W \mid \Phi\ N) \\ &\quad \propto P(\Phi \mid [A = \rho_A(m)])\ P(W \mid \Phi\ [N = \rho_\nu(m)]), \end{aligned} \tag{5.25}$$

where the summation over A and N are reduced to the terms $A = \rho_A(M)$ and $N = \rho_\nu(M)$ for which $P(A \mid M)$ and $P(N \mid M)$ are not zero (see Eqs (5.21) and (5.22)).

The result provided by Eq (5.25) states that selecting a control variable m is performed by combining two terms. The first term, $P(\Phi \mid [A = \rho_A(m)])$, corresponds to the confidence that the auditory consequence $\rho_A(m)$ would be categorized as the intended phoneme Φ; it implements the perceptual constraint in GEPPETO. The second term, $P(W \mid \Phi\ [N = \rho_\nu(m)])$, corresponds to the confidence that the predicted force $\rho_\nu(m)$ would satisfy the intended level of effort W, specified by phoneme Φ; it implements the effort constraint in GEPPETO. The resulting product of these two terms means that both probabilities need to be high in order to select control variables with high probability. In other words, high probability is given only to control variables that satisfy both constraints simultaneously.

3.3 Model implementation

The aim of the model is to generate motor control variables M performing a desired phoneme Φ with an intended level of effort W. The probability distribution $P(M \mid \Phi\ W)$, derived in Eq (5.25), characterizes every control variable M with its probability for achieving this task. One possible choice would be to select control variables that maximize this probability, however this would eliminate any possible variability and lead to the stereotyped situation encountered in GEPPETO. As our aim is to preserve variability, we adopt a decision policy based on a random sampling of the control variables space according to $P(M \mid \Phi\ W)$.

This sampling is implemented by a standard Markov Chain Monte Carlo algorithm (MCMC), which performs a random walk that converges to the desired distribution. Simulations were performed with Matlab's "`mhsample`" function with 20 chains of 2.10^4 samples each and a burning period of 10^3 samples. Tests revealed that further increasing the number of samples had no influence on the global shape of the obtained distributions.

We draw attention to the interpretation of this particular implementation of the Bayesian model. We are not assuming that biological systems are indeed performing MCMC sampling. In terms of a biological implementation of this process, one would imagine that the brain stores information about $P(M \mid \Phi\ W)$ in some way and would use it to optimize the mapping from phoneme and effort levels to motor space.

3.4 Results

We evaluate the results obtained by the inference process in four ways. We first analyze the distribution of motor variable M obtained under $P(M \mid \Phi\ W)$ for the different values of Φ and W. Next, in order to validate the performance of the model, we evaluate whether samples of control variable M effectively result in intended auditory and force values that satisfy both perceptual and effort constraints. Finally, for completeness, we present tongue configurations resulting from the inferred motor commands.

3.4.1 Distribution of control variable M inferred through $P(M \mid \Phi\ W)$

Since M is 6-dimensional, its distributions cannot be represented in a 2-dimensional figure. Instead, Fig 5.6 represents the histograms of samples for each of the components $(\lambda_1, \cdots, \lambda_6)$ obtained from $P(M \mid \Phi\ W)$ by the sampling process described in the previous section. These histograms approximate the marginals distributions of $P(M \mid \Phi\ W)$ along each of its dimensions.

It can be noted that each phoneme corresponds to a specific set of distributions of control variables λ. Some of these control variables appear to be constrained within small ranges of

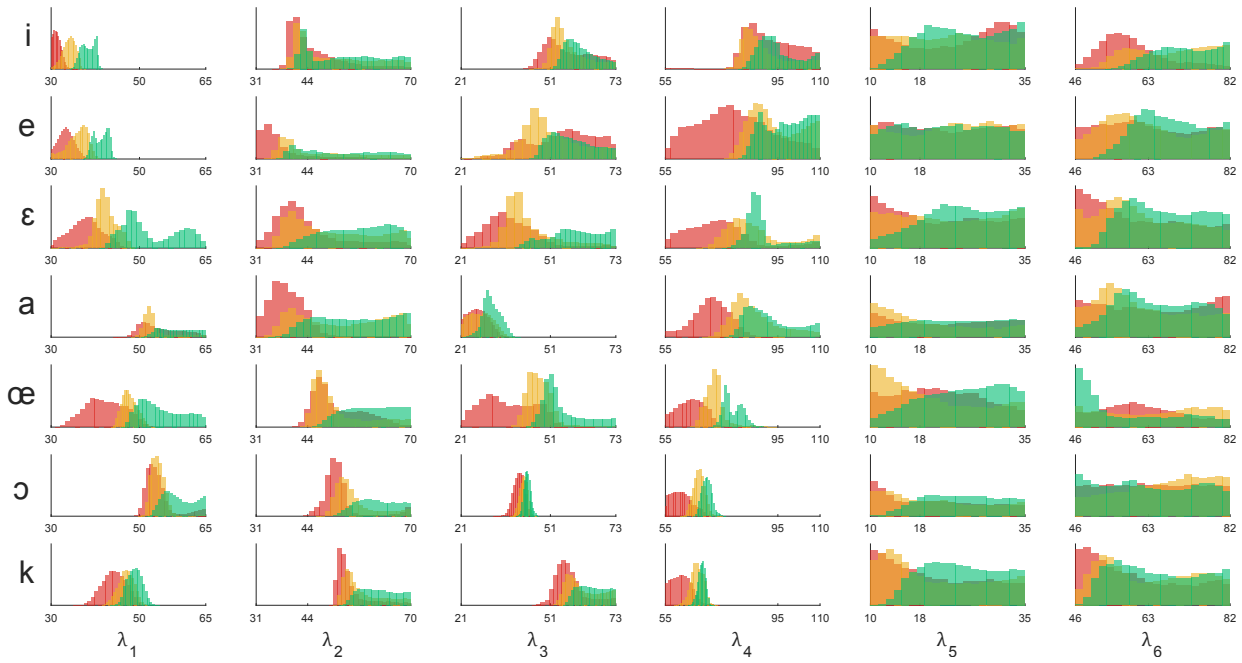

Figure 5.6: Histograms of the control variable samples obtained with the single phoneme planning model. 2.10^6 *samples of control variable* M *obtained from* $P(M \mid \Phi\ W)$*, for all phonemes and effort levels, with* $\kappa_A = 0.3$*. Rows correspond to phonemes, and columns to each* λ *component of the control variable* M*. Green histograms correspond to effort level* W *="Weak", yellow histograms correspond to effort level* W *="Medium" and red histograms correspond to effort level* W *="Strong". The corresponding muscles controlled by each control variable are:* λ_1*: Posterior Genioglossus,* λ_2*: Anterior Genioglossus,* λ_3*: Hyoglossus,* λ_4*: Styloglossus,* λ_5*: Verticalis,* λ_6*: Inferior Longitudinalis. Muscle lengths at rest (no generated force) are provided on each x-axis. Only* λ *parameters below these values may lead to active muscle force.*

values, for instance λ_3 in phoneme /ɔ/. Some other appear to have a wide range of variation, for instance λ_5 and λ_6 for all phonemes. This indicates varying roles of muscles in performing each phoneme. In addition we can see that planning with different levels of effort W leads to different distributions of λ values. Only values of λ_i that are smaller than the muscle rest length (indicated on the x-axes in Fig 5.6) lead to active muscle force, and the smaller the λ_i, the greater the force generated by muscle i. We observe that the lowest level of effort (green histograms) indeed always correspond to greater λ values than medium (yellow histograms) and strong (red histograms) effort levels.

In addition, we notice that control variables λ_1 and λ_3 negatively correlate for phonemes /i, e, ɛ, a/: smaller values of λ_1 are related to higher values for λ_3 and *vice versa*. The values taken by these control variables specify the activation level of the Posterior Genioglossus and Hyoglossus muscles. Small values of the λ control variables correspond to high levels of muscle activation and *vice versa*. We can thus see that the Bayesian model correctly extracts the antagonist interaction of these two muscles in the front-high and back-low movement direction. This antagonism has been found in electromyographic measures of muscle activity during speech production (K. Honda, 1996, Figure 2). This direction of movement is thus coherent with the variation of position of the tongue for the production of these four phonemes, qualitatively confirming the good adequacy of the Bayesian model with experimental results.

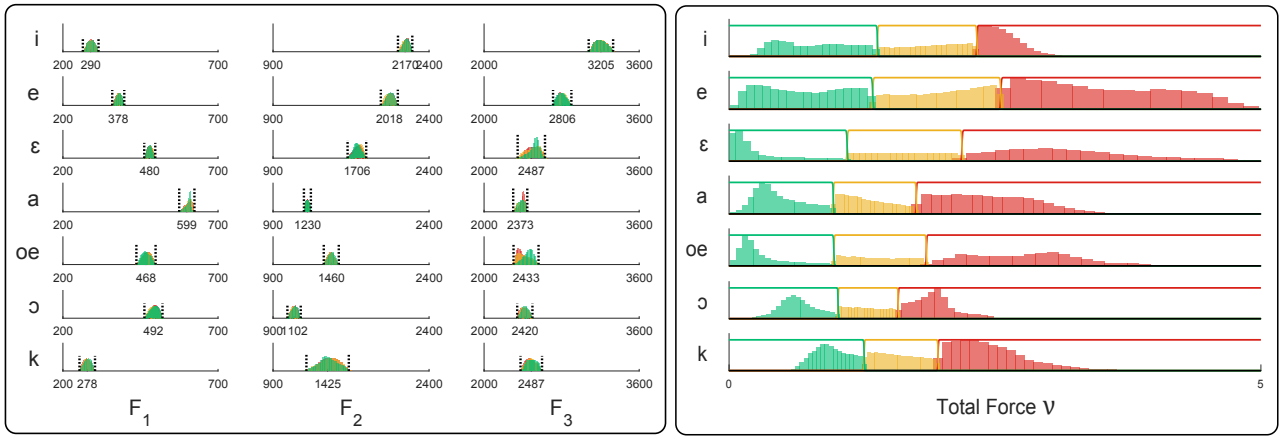

Figure 5.7: Histograms of auditory consequences (left panels) and forces (right panels) resulting from samples M obtained from the probability distribution $P(M \mid \Phi\ W\ [C_A = 1])$ (with $\kappa_A = 0.3$) for all phonemes and effort levels. *Green histograms correspond to effort level* W *="Weak", yellow histograms correspond to effort level* W *="Medium" and red histograms correspond to effort level* W *="Strong". The vertical dotted lines indicate borders of the auditory regions characterizing the auditory constraint for each phoneme; colored step-like curves on right panels indicate the corresponding regions characterizing each level of effort in GEPPETO.*

3.4.2 Satisfying the perceptual and effort constraints

We now aim to evaluate whether the samples of control variable M inferred through $P(M \mid \Phi\ W)$ effectively result in intended auditory and force values in order to satisfy both perceptual and effort constraints. Given the great number of samples, we use the RBF networks in order to assess the auditory and force consequences of each of the obtained values of control variable M.

Concerning the perceptual constraint, the left panel of Fig 5.7 represents the histograms of the first three formant values resulting from inferred samples M, for all phonemes and all effort levels. It can be observed that samples correctly distribute inside the auditory target regions for all phonemes and all effort levels.

Concerning the effort constraint, let us now evaluate whether these same samples of control variable M effectively generate total forces that distribute according to the intended effort levels. The right panel of Fig 5.7 represents the histograms of total muscle force corresponding to the same set of samples M used in previous Figures. It can be observed that the forces correctly distribute according to the corresponding regions that define each level of effort. In summary, the Bayesian model correctly infers control samples M that jointly satisfy both the auditory and the effort constraints characterizing the speech task.

3.4.3 Resulting tongue configurations

Finally, for illustration, Fig 5.8 represents the contour of 100 tongue configurations corresponding to control variables M obtained through $P(M \mid \Phi\ W)$ for each phoneme Φ and each level of effort W. These figures further indicate that planning under different levels of effort lead to similar tongue configurations.

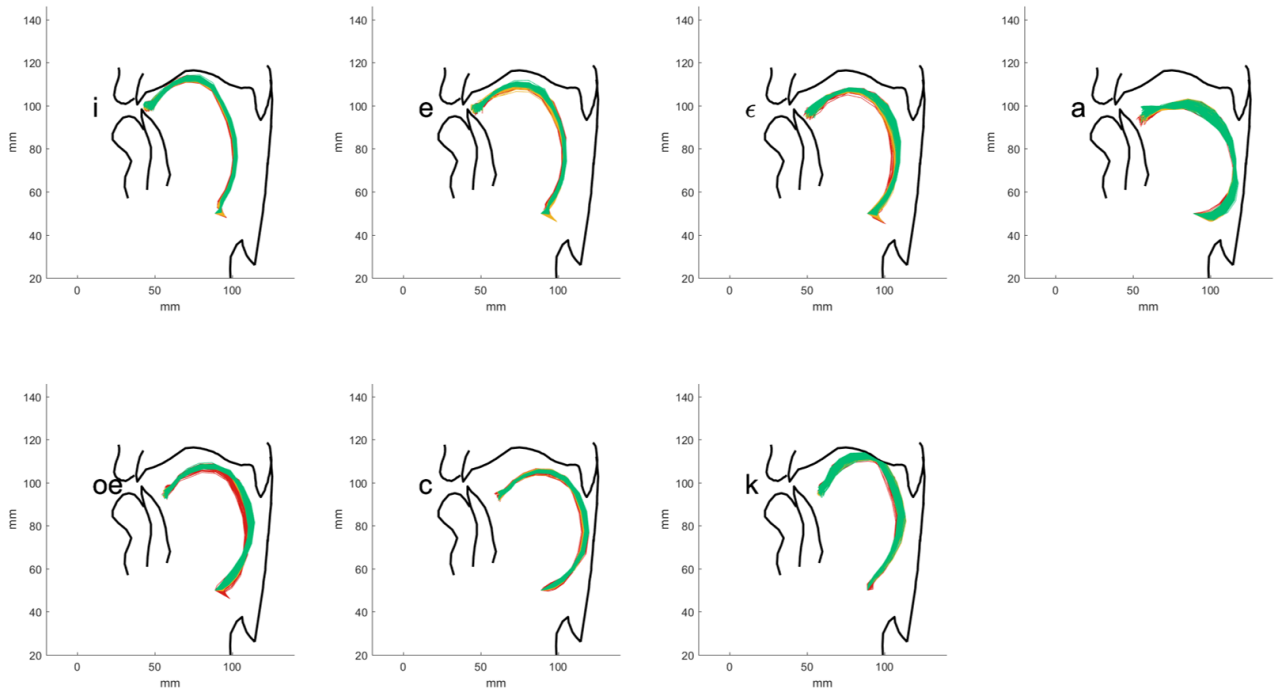

Figure 5.8: Tongue configurations resulting from samples M obtained from the the probability distribution $P(M \mid \Phi\ W)$ for different phonemes and different levels of effort. *Results are superimposed, with the same color code as in previous Figures. Different effort levels lead to similar tongue configurations, and thus curves hide each other in the plots.*

4 Bayesian model for planning a sequence of phonemes

The model presented in the previous section enables to infer control variables for the production of single phonemes. It would seem straightforward to use this same model for the production of a sequence of phonemes by just applying it for each phoneme in the sequence. However, this would result in a sequence of planning processes that are independent from each other, which would fail to reproduce the well known anticipatory coarticulation effects in particular. GEPPETO accounts for anticipatory coarticulation by assuming that sequences of control variables are not selected independently, but are constrained by a motor constraint that favors small displacements of control variables across the planned sequence. In this section, we build up on the previous single-phoneme model and extend it in order to obtain a planning process that involves the same motor constraint as GEPPETO. For simplicity, we will focus only in the case of three-phoneme sequences.

4.1 Model description

4.1.1 Variables

Planning a sequence of phonemes involves the same variables as the ones considered in the previous single-phoneme model. They correspond to the motor control variables M, auditory signals A, phonemes Φ, total forces N, and effort levels W. However, as we are considering a sequence instead of a single phoneme, each variable has to be repeated as many times as there are phonemes in the sequence. In order to distinguish each of the different instances of the variables, we label them with an index specifying their position in the sequence. Thus,

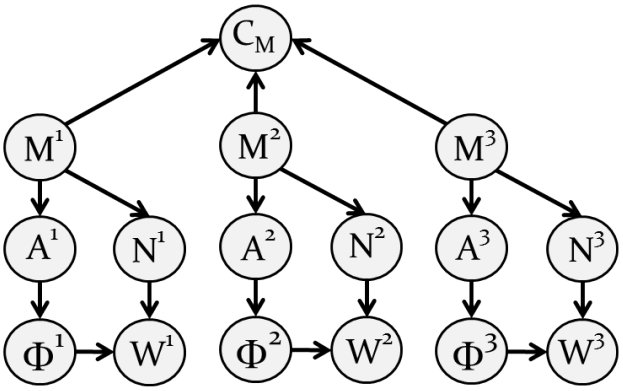

Figure 5.9: Graphical representation of the phoneme sequence model. *The diagram represents the decomposition of the joint probability distribution given by Eq (5.32). Nodes represent variables in the model and arrows represent their dependencies.*

variables become M^i, A^i, Φ^i, N^i and W^i, with $i \in \{1:3\}$, since we focus on three-phoneme sequences. For simplicity, we denote by $Y^{1:3} = \{Y^1 \wedge Y^2 \wedge Y^3\}$ the conjunction of different instances of a given variable Y at different positions in the sequence.

We also introduce an additional variable C_M in order to take into account the motor constraint assumed in GEPPETO. C_M is a binary variable that acts as a switch, activating the minimum motor displacement constraint when $C_M = 1$.

4.1.2 Decomposition

Let us denote by $X^i = \{M^i, A^i, \Phi^i, N^i, W^i\}$ the conjunction of all the variables at a given position i in the sequence. With this notation the joint probability distribution can be written as:

$$P(M^{1:3}\ A^{1:3}\ \Phi^{1:3}\ N^{1:3}\ W^{1:3}\ C_M) \quad = \quad P(X^{1:3}\ C_M). \tag{5.26}$$

Applying the product rule to the right hand side of Eq (5.26) leads to:

$$P(X^{1:3}\ C_M) \quad = \quad P(X^1)\ P(X^2 \mid X^1)\ P(X^3 \mid X^2\ X^1)\ P(C_M \mid X^3\ X^2\ X^1). \tag{5.27}$$

This expression can now be simplified thanks to the hypotheses made in GEPPETO. First, besides the minimum motor displacement constraint in GEPPETO, there is nothing creating any dependencies relating variables at different positions in the sequence. In real speech production, this type of constraint does exist, and corresponds to what is called "phonotactic" rules in linguistics, which are language dependent. It is not the purpose of the present study to address this type of high level linguistic constraints. We note though that the Bayesian Programming framework would enable to account for this kind of constraint.

The independence of variables at different positions in the sequence enable us to simplify the second and third factors in the decomposition of Eq (5.27) such that:

$$P(X^{1:3}\ C_M) \quad = \quad P(X^1)\ P(X^2)\ P(X^3)\ P(C_M \mid X^3\ X^2\ X^1). \tag{5.28}$$

The last factor in this decomposition corresponds to the dependence of the variable C_M on the other variables. The aim of C_M is to implement the motor constraint of GEPPETO. According to hypothesis $H_{6.3}$, the motor constraint only takes into account control variables $M^{1:3}$ at each

position in the sequence by computing the perimeter of the triangle defined by these control variables. Therefore, besides $M^{1:3}$, no other variable directly influences variable C_M and the last term in the decomposition of Eq (5.28) can be further simplified as:

$$P(C_M \mid X^3\ X^2\ X^1) \;=\; P(C_M \mid M^3\ M^2\ M^1). \tag{5.29}$$

Taking into account these simplifications, the joint probability distribution becomes:

$$P(X^{1:3}\ C_M) \;=\; P(X^1)\ P(X^2)\ P(X^3)\ P(C_M \mid M^3\ M^2\ M^1). \tag{5.30}$$

From this last expression it can be seen that the decomposition of the joint probability distribution $P(X^{1:3}\ C_M)$ contains three copies of the probability distribution $P(X^i)$. We identify each of these terms to the joint probability distribution of variables involved in the production of single phonemes, as derived in Section 3:

$$P(X^i) \;=\; P(M^i)\ P(A^i \mid M^i)\ P(\Phi^i \mid A^i)\ P(N^i \mid M^i)\ P(W^i \mid \Phi^i\ N^i). \tag{5.31}$$

Combining Eq (5.30) with Eq (5.31) gives the complete decomposition of the joint probability distribution:

$$\begin{aligned} &P(M^{1:3}\ A^{1:3}\ \Phi^{1:3}\ N^{1:3}\ W^{1:3}\ C_M) \\ &= P(M^1)\ P(A^1 \mid M^1)\ P(\Phi^1 \mid A^1)\ P(N^1 \mid M^1)\ P(W^1 \mid \Phi^1\ N^1) \\ &\quad P(M^2)\ P(A^2 \mid M^2)\ P(\Phi^2 \mid A^2)\ P(N^2 \mid M^2)\ P(W^2 \mid \Phi^2\ N^2) \\ &\quad P(M^3)\ P(A^3 \mid M^3)\ P(\Phi^3 \mid A^3)\ P(N^3 \mid M^3)\ P(W^3 \mid \Phi^3\ N^3) \\ &\quad P(C_M \mid M^3\ M^2\ M^1). \end{aligned} \tag{5.32}$$

Fig 5.9 represents the Bayesian network corresponding to the decomposition given by Eq (5.32).

4.1.3 Parametric forms

Having derived the decomposition of the joint probability distribution, we now determine the form taken by each of the factors in Eq (5.32). This was already done in Section 3 for the terms appearing in the first three lines in Eq (5.32). The last term, $P(C_M \mid M^3\ M^2\ M^1)$, represents the dependence of variable C_M on the control variables. The aim of the motor constraint in GEPPETO is to penalize patterns of control variables that are far from each other by attributing them a cost that increases with the perimeter of the triangle that they define in the control space ($H_{6.3}$). The same motor constraint is implemented in $P(C_M \mid M^3\ M^2\ M^1)$ through:

$$P\left([C_M = 1] \mid M^3\ M^2\ M^1\right) \;:=\; e^{-\kappa_M\left(|M^2-M^1|+|M^2-M^3|+|M^3-M^1|\right)}. \tag{5.33}$$

The additional parameter κ_M is introduced in order to modulate the strength of this motor constraint. The motor constraint defined by Eq (5.33) is interpreted in the following way. The further the control variables are from each other, the smaller the probability for the variable C_M to be in state $C_M = 1$. Therefore, when state C_M is assumed to be 1, motor control variables that are close to each other are more probable.

For completeness, as C_M takes only two values, the corresponding expression for the probability of having $C_M = 0$ is given by:

$$\begin{aligned} P([C_M = 0] \mid M^3\ M^2\ M^1) &= 1 - P([C_M = 0] \mid M^3\ M^2\ M^1) \\ &= 1 - e^{-\kappa_M\left(|M^2-M^1|+|M^2-M^3|+|M^3-M^1|\right)}, \end{aligned} \tag{5.34}$$

which would correspond to a maximum motor displacement constraint.

4.2 Question: Motor planning in the context of a sequence of phonemes

Considering the planning problem addressed in GEPPETO, the task assigned to the present model is to infer a sequence of motor control variables $M^{1:3}$ under the condition that the desired phonemic categories $\Phi^{1:3}$ are reached with the intended level of effort and imposing the activation of the motor constraint with $C_M = 1$. This inference is formulated in Bayesian terms by $P(M^{1:3} | \Phi^{1:3}\ W^{1:3}\ [C_M = 1])$. This is again solved in a standard way through the knowledge provided by the joint probability distribution of Eq (5.32). The corresponding expression is given by:

$$\begin{aligned}
&P\left(M^{1:3} \mid \Phi^{1:3}\ W^{1:3}\ [C_M = 1]\right) \\
&\quad \propto \sum_{\{A^{1:3}, N^{1:3}\}} P(X^1)P(X^2)P(X^3)P\left([C_M = 1] \mid M^3M^2M^1\right) \\
&\quad \propto \ P\left([C_M = 1] \mid M^3M^2M^1\right) \sum_{A^{1:3}, N^{1:3}} \prod_{i=1}^{3} P(X^i) \\
&\quad \propto \ P\left([C_M = 1] \mid M^3M^2M^1\right) \prod_{i=1}^{3} \sum_{A^i, N^i} P(M^i\ A^i\ \Phi^i\ N^i\ W^i) \\
&\quad \propto \ P\left([C_M = 1] \mid M^3M^2M^1\right) \prod_{i=1}^{3} P\left(\Phi^i \mid \rho_a(M^i)\right) P\left(W^i \mid \Phi^i\ \rho_\nu(M^i)\right) \qquad (5.35)
\end{aligned}$$

where the proportionality symbols account for normalization constants. The last line results from the same observation as in Section 3 for the sums over A^i and N^i of the joint probability distribution $P(M^i\ A^i\ \Phi^i\ N^i\ W^i)$.

4.3 Results

The Bayesian three-phoneme model is implemented through Monte Carlo sampling as described in Section 3.3. The performance of the previous single-phoneme model was evaluated in relation to its capacity to produce acoustic outputs and total forces that satisfy the perceptual and effort constraints characterizing the task. For the present three-phoneme model, performance is further evaluated with respect to the minimum displacement constraint imposed to the sequence of selected control variables (M^1, M^2, M^3).

For simplicity, we focus on motor planning that do not specify particular values of the effort constraint. In this particular case Eq (5.35) becomes:

$$\begin{aligned}
&P\left(M^{1:3} \mid \Phi^{1:3}\ [C_M = 1]\right) \\
&\quad \propto \ P\left([C_M = 1] \mid M^3M^2M^1\right) \prod_{i=1}^{3} P\left(\Phi^i \mid \rho_a(M^i)\right) \qquad (5.36)
\end{aligned}$$

4.3.1 Perceptual constraint: Acoustic outputs and coarticulation

Concerning the perceptual constraint, Fig 5.10 illustrates the projection onto the (F_1, F_2) plane of the acoustic consequences of 100 inferred motor control variables (M^1, M^2, M^3) for the production of the sequence /aki/. The optimal solution obtained under similar conditions with GEPPETO is also displayed for comparison.

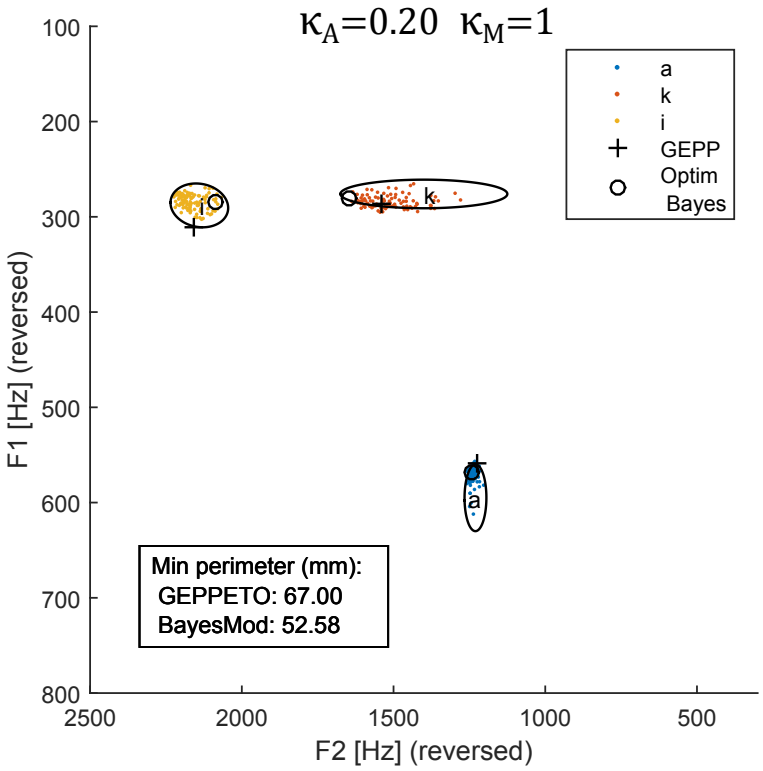

Figure 5.10: Projection, on the (F_2, F_1) plane, of the auditory outputs resulting from motor planning in the phoneme sequence model. *Auditory outputs are obtained from 100 samples of the control variable M obtained from the inferred probability distribution for the production of sequence /aki/. The auditory output obtained by GEPPETO and the sample of the Bayesian three-phoneme model with minimum perimeter are also indicated. Values of the perimeters obtained by each model are indicated in the legend.*

The first thing to notice is the variability of the results obtained from the Bayesian three-phoneme model. This was expected given the probabilistic framework of the model. Secondly, it can be observed that the obtained spectral patterns effectively distribute inside the correct target regions. This illustrates that the three-phoneme model correctly satisfies the perceptual constraint. Thirdly, point clouds characterizing the distributions of the resulting spectral properties are shifted from the center of the target regions toward their boundaries, with a clear tendency for the /a/ productions to be shifted to smaller F_1 values, and for the /k/ productions to be shifted toward higher F_2 values. This shows the influence of the motor constraint on the planned sequence at the acoustic level. This is also observed in the sounds obtained with GEPPETO.

Finally, Fig 5.11 illustrates the role of parameters κ_M and κ_A in the fulfillment of the perceptual constraint. These parameters are related to precision (or inverse variance) modulating the certainty, or confidence, of the motor and perceptual constraints in the model. The stronger the weight of the motor constraint (κ_M), relative to the perceptual constraint (κ_A), the stronger the shift of the points from the center of the phoneme target regions and towards the center of the vocalic triangle, that would result from the neutral motor command. At the extreme, targets are no longer reached if the value of κ_M becomes too large compared to κ_A as can bee seen in the center and right panels of Figure 5.11.

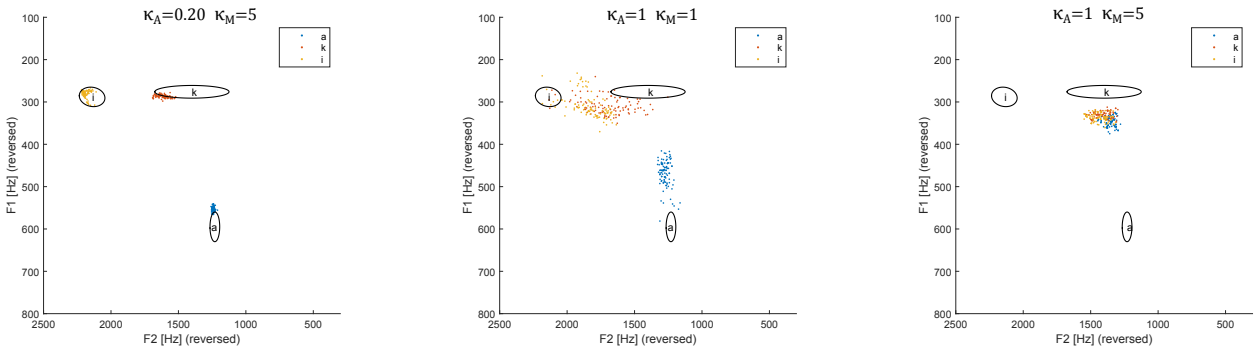

Figure 5.11: Effect of parameters κ_A and κ_M on the spectral properties of the acoustic signals obtained by the Bayesian three-phoneme model. *Refer to Figure 5.10 for comparison. Left: Keeping κ_A to the same value as in Figure 5.10 and augmenting κ_M by a factor 5. Middle: Keeping κ_M to the same value as in Figure 5.10 and multiplying κ_A by a factor 5 (remember that augmenting κ_A corresponds to relaxing the constraint, see Figure 5.2). Right: Augmenting the motor constraint and relaxing the perceptual constraint at the same time. Phonemic targets are attained as long as there is a correct balance between the strength of the motor constraint and the strength of the perceptual constraint.*

4.3.2 Motor constraint

We have seen the effect of the motor constraint on the planned sequence at the acoustic level: acoustic signals deviate from the center of the target regions and tend to be closer from each other. However, it should be noted that the minimization of the motor cost occurs in the motor space and not in the acoustic space. Hence, the closer proximity of spectral realizations of the phonemes in the sequence is a consequence in the acoustic space of the constraint in the motor space. This explains in particular that, in the left panel of Fig 5.11, the spectral characteristics of the selected realizations of phoneme /i/ appear to deviate away from the two other phonemes, instead of being close to them, as one would expect from the form of the motor constraint. A tentative explanation for this phenomenon is the strong non-linearity of the mapping relating motor control variables to the spectral properties of the acoustic signal, observed in particular for vowel /i/.

In order to evaluate whether the motor constraint actually performs the minimization of the distance between motor control variables involved in the sequence, it is necessary to evaluate the actual perimeter of the triangle that they define in the motor space. Fig 5.12 shows the average value taken by this perimeter for 100 inferences of the Bayesian three-phoneme model for the sequence /aki/, as a function of the parameter κ_M and for different values of κ_A. The value obtained with GEPPETO is also presented for comparison.

We first note that curves corresponding to different values of κ_A all merge for $\kappa_M = 0$. This corresponds to the situation where there is no motor constraint, and therefore planning the sequence is performed independently for each phoneme of the sequence. The average value of the distance between control variables does not depend on κ_A in that case. Next, we observe that the average perimeter is clearly reduced when the strength of the motor constraint is raised with κ_M, and the capacity to minimize the motor cost is stronger for higher values of κ_A (i.e., for weaker perceptual constraints, see Figure 5.2). This illustrates the trade-off between the two constraints governed by κ_M and κ_A in the Bayesian three-phoneme model.

Finally, we evaluate the way the Bayesian three-phoneme model performs compared with GEPPETO. It can be noted in Figure 5.12 that, for each value of κ_A, (i.e., each level of percep-

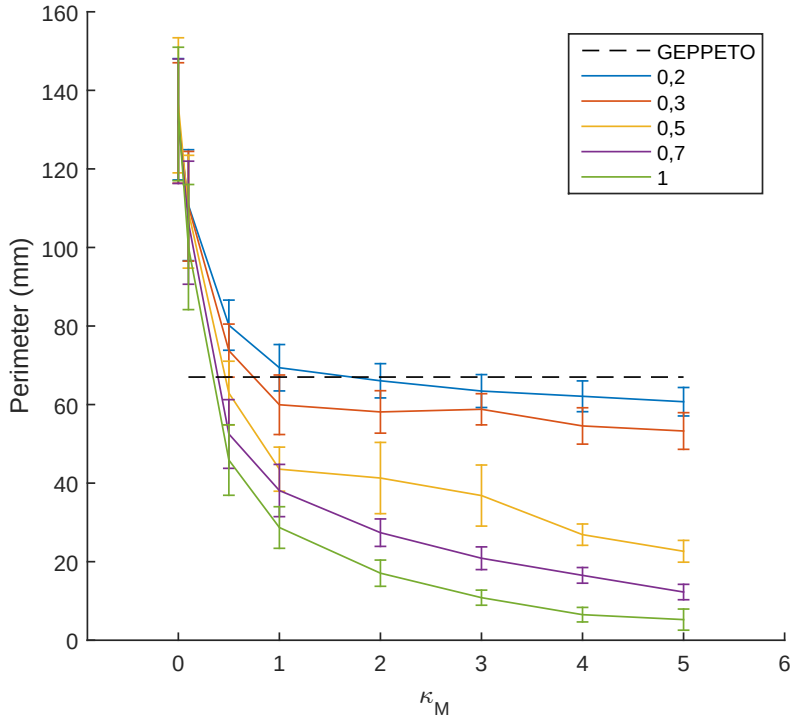

Figure 5.12: Average distances obtained with the Bayesian three-phoneme model as a function of parameter κ_M. *Results for different values of κ_A are plotted (listed in the insert on the right). Error bars indicate variability obtained over 100 random samples. The black horizontal dashed line represents the value obtained with GEPPETO.*

tual constraint) there is a value of κ_M (i.e., a strength of the motor constraint) for which the average distance between control variables obtained with the Bayesian three-phoneme model coincides with the result obtained with GEPPETO. For instance, for $\kappa_M = 1$, the Bayesian three-phoneme model coincides with GEPPETO when $\kappa_A = 0.2$. Figure 5.10 confirms that for these specific parameter values, the perceptual constraint is correctly satisfied and the spectral characteristics obtained with the Bayesian three-phoneme model are close to the spectral characteristics obtained with GEPPETO. This suggests an equivalence of the two models for these specific values of the parameters. However, if we compare the optimal control variables obtained with the Bayesian three-phoneme model, i.e., those that minimize the perimeter in the motor control space, with the optimal commands obtained with GEPPETO, we observe that the optimal perimeter obtained by the Bayesian three-phoneme model is actually smaller than the one obtained by GEPPETO. This suggests that GEPPETO has not found the true optimal values. We will return to this issue in the next section.

4.3.3 Speaking rates

We end this section with an illustration of the interest of the effort constraint for achieving accuracy during fast speaking rates. We consider the sequence /aiɛ/ planned with the W="Weak" and W="Strong" levels of effort. The set of control variables having the highest inferred probabilities are selected and the resulting tongue trajectories are generated by the biomechanical model as presented in Chapter 4.

Two speaking rates are implemented for each set of control variables by specifying a slow

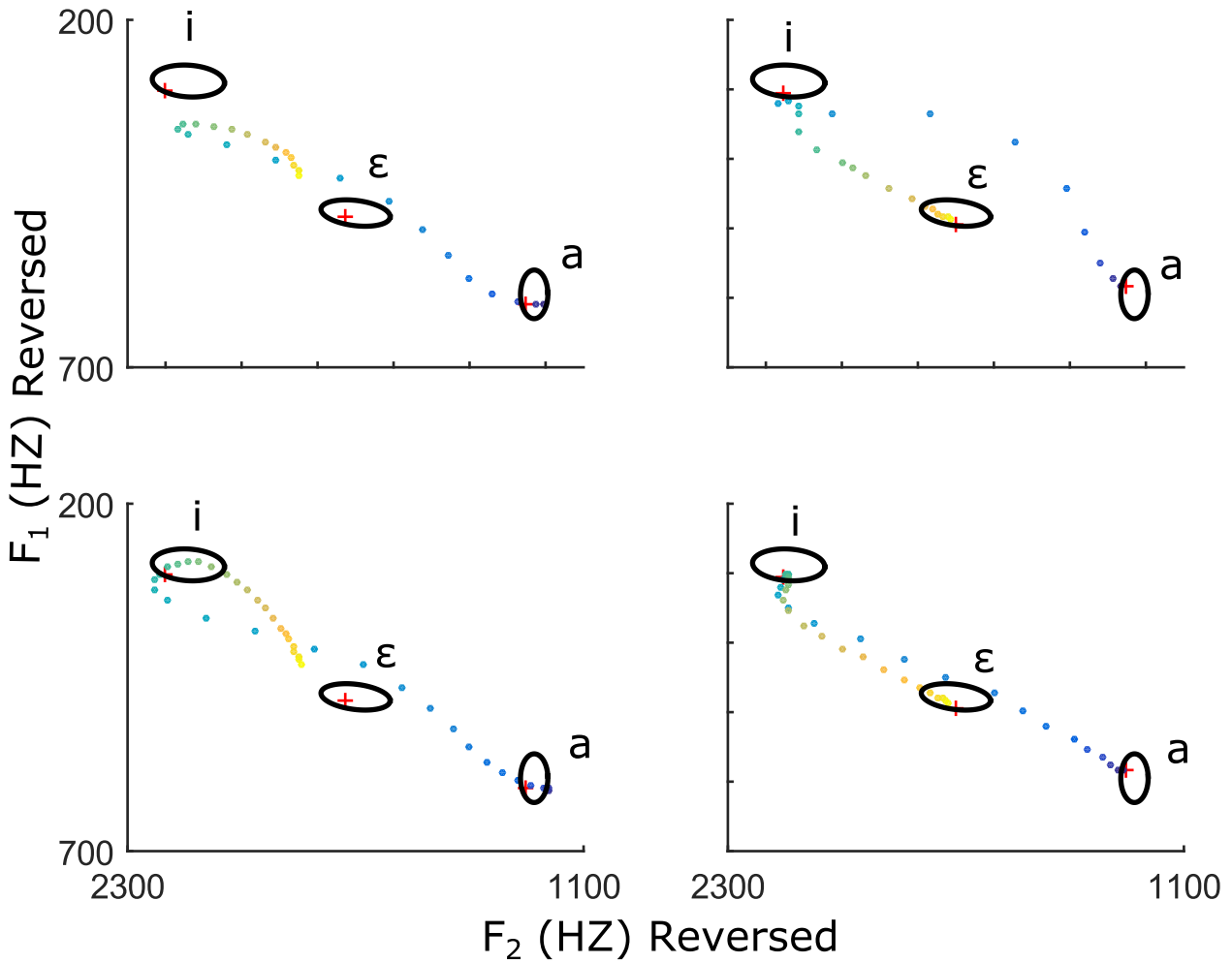

Figure 5.13: Formant trajectories corresponding to the control variables planned for the sequence /aiɛ/ with the W="Weak" (left panels) and W="Strong" (right panels) effort levels, produced with fast (Top panels) and slow (bottom panels) speech rates. *For a fast speaking rate, the trajectory planned with the W="Weak" effort level misses the auditory target for the middle phoneme /i/ (Top left), but not for the W="Strong" effort level (Top right). Both commands produce trajectories that reach the auditory targets for slow speech rate (Bottom). Red crosses indicate the intended auditory output of the control variables.*

and fast transition rate between the control variables of the first and second phoneme in the sequence, together with a long and a short duration of the second phoneme. Results are illustrated as formant trajectories in auditory space shown in Fig 5.13. It can be seen that for a fast speaking rate, control variables planned with the W="Weak" level of effort result in a formant trajectory that misses the auditory target region for the middle phoneme /i/ (top left image). This is not the case for control variables planned with the W="Strong" levels of effort (top right image). In a slow speaking rate however, these same two sets of control variables result in formant trajectories that both reach auditory target regions (bottom images).

5 Discussion

5.1 Model equivalence

We have described a Bayesian reformulation of the motor planning stage of the GEPPETO model, initially formulated with an optimal control approach. Simulation results indicate that, as for its optimal control version, the Bayesian three-phoneme model correctly infers motor control variables that perform the desired motor task satisfying the specified perceptual, effort and motor constraints. Furthermore, for specific values of the parameters κ_A and κ_M characterizing the strengths of the perceptual and motor constraints, simulations suggest the equivalence of results obtained by both models. This equivalence is evaluated on the basis of the comparison between average values obtained with the Bayesian model and the optimal solution obtained with GEPPETO.

Nevertheless, it can be shown that the optimal control model can be obtained as a particular case of the Bayesian three-phoneme model if one looks for the configuration of control variables that maximize the posterior probability given by $P(M^{1:3} \mid \Phi^{1:3}\ W^{1:3})$. The derivation of this result is provided as a supplementary material (see Annex A) and rests on the property that the negative logarithm of $P(M^{1:3} \mid \Phi^{1:3}\ W^{1:3})$ turns out to be equivalent to the cost function of GEPPETO. Therefore, maximizing the probability $P(M^{1:3} \mid \Phi^{1:3}\ W^{1:3})$ is identical to minimizing the equivalent cost function of GEPPETO, showing that the Bayesian three-phoneme model can be simplified to GEPPETO in this specific implementation scheme.

Note that there are mathematical theorems showing that a Bayesian scheme exists for any set of cost functions and optimal behavior. These are known as complete class theorems (Brown, 1981; Robert, 2007). Knowing this theoretical context, stating that the Bayesian reformulation of GEPPETO is able to account for its optimal control scheme is not surprising. However, the theorems state the existence of the Bayesian reformulation; our contribution goes further, by defining the structure taken by this reformulation in our case. This is discussed in more detail in the sections below.

5.2 Addressing redundancy and variability in formal terms

We were interested in the problem of how a feedforward model of motor planning can solve the indeterminacy characterizing the specification of motor control variables for achieving a desired motor task, without resulting in a stereotyped behavior. The essence of the dilemma was rooted in the fact that, on the one hand, indeterminacy arises from redundancy, i.e., from the multiplicity of solutions to the problem, and on the other hand, solving redundancy, i.e., eliminating all possible solutions but one, inevitably results in stereotypy. We suggested that variability could be recovered at this point by assuming that even if the planning problem is driven by an optimality assumption, the actual solution might not be a stereotyped one.

The absence of stereotypy may be first due to inherent computational limitations of the search for optimal solutions. In GEPPETO, the optimization algorithm relies on a gradient descent scheme. Crucially, due to non-linearities between variables of the model, the cost function may feature multiple local minima and the solutions obtained by gradient descent techniques may be highly dependent on the initial values of the optimization algorithm. Initializing the gradient descent algorithm in GEPPETO with different starting positions does indeed drive convergence to different locally "optimal" solutions. In particular this explains why the solution obtained with GEPPETO, as shown in Figure 5.10, appears to have a greater perimeter value than the optimal solution found with the Bayesian three-phoneme model. The result for GEP-

PETO was actually chosen as the best one out of 100 different initializations of the descent algorithm. The fact that the gradient descent algorithm has failed to converge to solutions as good as those found by its Bayesian reformulation in all of these 100 initializations indicates the degree of complexity of the optimization process.

In this context, it could be argued that variability in speech production arises from the existence of these multiple local solutions towards which the optimization process may differently converge, depending on its initial configuration. However, the variability introduced this way cannot be formally justified as actually arising from the model itself, since it is just an indirect consequence of the failure of its implementation for finding the true optimal solution that the model actually predicts. Moreover, this *ad hoc* implementation can only account for variability in a qualitative way and does not have any theoretical or cognitive foundation.

In contrast, formulating the feedforward planning process within a Bayesian modeling framework has allowed us to address the indeterminacy of the problem and deal with variability in formal terms. This is made possible by the fact that the Bayesian approach does not solve indeterminacy by suppressing all solutions but one. Instead, the Bayesian framework characterizes every possible configuration by its probability to achieve the task. Redundancy is then solved by our decision process, in which we randomly select motor control variables under the corresponding probability distribution. Variability becomes therefore an inherent consequence of the formalism and the chosen decision process. Furthermore, the variability generated with this approach has a specific structure that relates to the ways knowledge is assumed to be represented in the brain.

Therefore, the advantage of the Bayesian modeling approach is to suggest that a probabilistic description of the planning process is able to deal with the selection of solutions to an ill-posed problem without destroying variability (Colas, Diard, & Bessière, 2010). This allows to treat variability in formal terms and not as the result of an *ad hoc* implementation of the model.

The pertinence of an approach that designs models integrating multiple local solutions in formal terms is illustrated by the work of Ganesh, Haruno, Kawato, and Burdet (2010). Their work indicates that motor memory plays a crucial role influencing the outcome of the planning process, in addition to the optimization of cost related to error and effort. Thus, motor memory would be responsible for setting the initial states of variables of the motor system, which would influence the convergence of the search for optimal solutions toward local optima. Even if the Bayesian model that we have presented does not account for the role of motor memory in the planning process, the Bayesian modeling approach offers a framework in which motor memory could be modeled via a set of local approximations to the complete probability distribution, as it would be performed by local Laplace approximation or by standard variational inference methods. This raises the question of how agents would encode the knowledge described by the probability distributions involved in the presented scheme. While a complete representation of a complex knowledge would involve an important amount of resources, it would be natural to select a simpler approximation to this knowledge as it would be advantageous for the agent and often sufficient for practical purposes.

Chapter 6

Alternative formulation of speech motor goals

1 Introduction

One of the fundamental hypotheses of GEPPETO is that phonemic motor goals are defined according to categorical perception. In other words, all auditory values inside the intended phoneme regions are assumed to be equivalent for planning. A consequence of this is that the distribution of productions for isolated phonemes is expected to be close to a uniform distribution inside the intended target regions. This contrasts with the common assumption that the distribution of productions of isolated vowels would follow Gaussian distributions.

In this section we reexamine the perceptual constraint specifying motor goals in GEPPETO and suggest an alternative formulation that does not rest on categorical perception. For simplicity, we focus the discussion on the particular case of single phonemes. We begin by introducing the general principle of this alternative approach in Section 2. Then, in Section 3, we implement this approach for the construction of a model for motor planning that does not involve categorical perception. Finally, in Section 4, we briefly discuss the implications of these two alternative approaches.

2 Avoiding the categorical perception module – coherence variables

The Bayesian model presented in the previous chapter performed motor planning by combining two main components: (1) the auditory-motor knowledge, $P(A \mid M)$, corresponding to the internal forward model relating control variables M to their auditory consequences A; and (2) the auditory categorization process $P(\Phi \mid A)$ corresponding to the perceptual constraint in GEPPETO. This categorization process was defined as the outcome of a sub-model, $P(\Phi_c \mid A_c)$, that was based on the auditory characterization of phonemes, $P(A_c \mid \Phi_c)$, identified with the expected distribution of auditory values corresponding to each phoneme. Our aim here is to define a variant model, based on the same set of knowledge, $P(A \mid M)$ and $P(A_c \mid \Phi_c)$, but without the involvement of the sub-module for the auditory categorization process.

To do so, an intuitive idea would be to identify variables A and A_c on the one hand, and variables Φ_c and Φ on the other hand, in order to write a decomposition of the joint probability distribution $P(M\ A\ \Phi)$ as:

$$P(M\ A\ \Phi) \quad \propto \quad P(A \mid M)\ P(A \mid \Phi)\ P(\Phi)\ P(M). \tag{6.1}$$

However, this would involve two probability distributions over the same variable A, which is forbidden by the product rule of Eq (5.6).

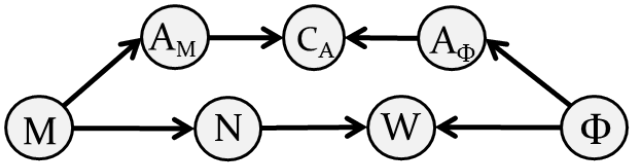

Figure 6.1: Graphical representation of the single phoneme model formulated with coherence variables. *The diagram represents the decomposition of the joint probability distribution given by Eq (6.3).*

This issue can be overcome with a mathematical trick that consists in keeping variable A duplicated while forcing its two instances to behave identically. More precisely, let us denote A_M the auditory predictions resulting from the auditory-motor internal model $P(A_M \mid M)$, and A_Φ the auditory expectations associated with the characterization of phonemes $P(A_\Phi \mid \Phi)$. In order to force these two variables to behave identically, we include an additional term imposing a "coherence constraint" between them (Bessière et al., 2013; Gilet, Diard, & Bessière, 2011). This coherence constraint is implemented by including an additional binary variable C_A, which we call a "coherence variable", that imposes the coherence between variables A_M and A_Φ when $C_A = 1$ with:

$$P([C_A = 1] \mid [A_M = a_1]\ [A_\Phi = a_2]) \quad = \quad \begin{cases} 1 & \text{if } a_1 = a_2 \\ 0 & \text{otherwise.} \end{cases} \tag{6.2}$$

3 Motor planning without categorical perception

In this section we present the alternative formulation of motor planning that exploits coherence variables in order to avoid the involvement of the categorical perception sub-module. We begin by describing the definition of the model in Section 3.1, we derive the inference question for motor planning in Section 3.1.2 and then present its results in Section 3.2.

3.1 Model definition

This model aims at implementing the same hypotheses involved in the model presented in Chapter 5, but following the coherence variable approach described in Section 2. The definition of variables, the decomposition of their joint probability distribution and the identification of terms in this decomposition are thus motivated in the same set of knowledge as in Chapter 5.

3.1.1 Description

Variables Control variables, phonemes, total force and effort levels remain represented by the same probabilistic variables M and Φ N and W as in Chapter 5. However, auditory signals are associated with two auditory variables, A_M and A_Φ, representing respectively auditory-motor predictions and auditory-phonological expectations. Finally, a coherence variable C_A implements the connection between the two auditory variables A_M and A_Φ, as described in Section 2.

Decomposition We decompose the joint probability distribution of all variables in the model in the following way:

$$\begin{aligned} P(M\ \Phi\ A_M\ A_\Phi\ C_A\ N\ W) &= P(M)\ P(\Phi)\ P(A_M \mid M)\ P(A_\Phi \mid \Phi) \\ &\quad P(C_A \mid A_\Phi\ A_M) \\ &\quad P(N \mid M)\ P(W \mid \Phi\ N). \end{aligned} \tag{6.3}$$

This decomposition, illustrated in Fig 6.1, involves a set of conditional independence assumptions that we do not justify here, but are similar to the ones discussed in Chapter 5. The first line in this Eq (6.3) corresponds to the relation intended in Eq (6.1); the second line corresponds to the addition of the coherence constraint term; the last line corresponds to the same decomposition involving force and effort constraint as in Chapter 5.

Parametric forms The identification of terms in Eq (6.3) follow the same hypotheses as terms involved in Chapter 5. No prior knowledge is assumed for M and Φ and thus both $P(M)$ and $P(\Phi)$ are identified with uniform probability distributions. $P(A_M \mid M)$ and $P(N \mid M)$ are identified with the motor-to-auditory and motor-to-force internal models of GEPPETO and are defined as Dirac probability distributions as in Chapter 5. $P(A_\Phi \mid \Phi)$ is identified with the auditory characterization of phonemes, defined in terms of Gaussian probability distributions as in Chapter 5. $P(W \mid \Phi\ N)$ is identified with the effort constraint of GEPPETO as defined in Chapter 5. Finally, $P(C_A \mid A_\Phi\ A_M)$ implements the coherence constraint as defined in Eq (6.2).

3.1.2 Question: Motor planning

In order to introduce how motor planning is formulated in this new model, we begin by focusing only in the inference of the motor variable M for the production of an intended phoneme Φ, without considering the specification of an intended level of effort. In this context variables M and Φ appear to be independent unless the coherence constraint is activated. Indeed, inferring motor commands as $P(M \mid \Phi)$ gives $P(M \mid \Phi) = P(M)$, which is the uninformed prior distribution over motor commands.

Therefore, in order to infer values of the control variable that achieve the expected auditory distribution of the intended phoneme, auditory variables A_M and A_Φ must be connected by imposing that the coherence constraint is activated ($C_A = 1$). The resulting motor planning process is therefore formulated as $P(M \mid \Phi\ [C_A = 1])$, and computing this inference from the model gives:

$$\begin{aligned} P([M = m] \mid \Phi\ [C_A = 1]) &\propto \sum_{A_M, A_\Phi, N, W} P([M = m]\ \Phi\ A_M\ A_\Phi\ [C_A = 1]\ N\ W) \\ &\propto \sum_{A_M, A_\Phi} P(A_M \mid [M = m]) P(A_\Phi \mid \Phi)\ P([C_A = 1] \mid A_\Phi\ A_M) \\ &\propto \sum_{A_\Phi} P(A_\Phi \mid \Phi)\ P([C_A = 1] \mid A_\Phi\ [A_M = \rho_a(M)]) \\ &\propto P([A_\Phi = \rho_a(m)] \mid \Phi), \end{aligned} \tag{6.4}$$

where summation terms involving variables N and W on the first line simplify to one. Furthermore, the summation over A_M on the second line reduces to the term $A_M = \rho_A(M)$ for which $P(A \mid M)$ is not zero and the summation over A_Φ on the third line reduces to the term $A_\Phi = \rho_A(M)$ for which $P(C_A \mid A_\Phi\ [A_M = \rho_a(M)])$ is not zero. Inferring values of the control

variable M for the production of an intended phoneme Φ with an intended level of effort W is then defined in a similar way as $P(M \mid \Phi\ [C_A = 1]\ W)$. Computing this inference from the model gives:

$$\begin{aligned} P([M = m] \mid \Phi\ [C_A = 1]\ W) &\propto \sum_{A_M, A_\Phi, N} P([M = m]\ \Phi\ A_M\ A_\Phi\ [C_A = 1]\ N\ W) \\ &\propto P([A_\Phi = \rho_a(m)] \mid \Phi)\ P(W \mid \Phi\ [N = \rho_\nu(m)]). \end{aligned} \quad (6.5)$$

3.2 Results

The aim of this section is to evaluate the performances of this new model by comparing it with the results obtained by the previous model presented in Chapter 5. We begin by focusing only on motor planning for the production of an intended phoneme Φ, without considering the specification of an intended level of effort, i.e., $P([M = m] \mid \Phi\ [C_A = 1])$ as given by Eq (6.4). Then, in Section 3.2.2, we evaluate results of $P([M = m] \mid \Phi\ [C_A = 1]\ W)$, as provided by Eq (6.5), for all phonemes and effort levels.

3.2.1 Comparison of motor planning with an without categorical perception

Eq (6.4) states that the motor planning process obtained with $P([M = m] \mid \Phi\ [C_A = 1])$ infers values m of the control variable M by evaluating their auditory consequences, $\rho_a(m)$, with respect to the expected distribution characterizing the intended phoneme, $P([A_\Phi = \rho_a(m)] \mid \Phi)$. Since we assumed these expected distributions to be Gaussian, control strategies that lead to acoustic signal that are closer to the center of the intended target region are more likely to be produced than those that are more distant.

This has to be contrasted with the motor planning process obtained by the model in Chapter 5, where values of the control variables M were inferred by evaluating their auditory consequences $\rho_a(m)$ with respect to the auditory categorization process defined by the sub-module $P(\Phi \mid A)$. As previously noted, this previous motor planning process rates equivalently all control variables as long as their resulting auditory outputs correspond to the plateau regions of $P(\Phi \mid A)$ (see Fig 5.2, right panel). In particular, control strategies that lead to acoustic signals that lie right in the center of the intended target region are as good as those that lie close to its borders.

Therefore, the motor planning process involving categorical perception predicts that the distribution of planned productions of isolated phonemes should be close to uniform inside the corresponding categorical regions. Note however that, since the mapping between motor commands to auditory outputs is many-to-one, uniformity is exact only if each auditory output has the same number of inverse images in motor space. This may not be true in practice; however it seems reasonable to assume that these differences are small and distributed homogeneously, such that the overall distribution remains close to uniform.

Fig 6.2 illustrates these differences by comparing histograms of auditory samples obtained from both planning processes for each phoneme. A difference between planning process can be observed: histograms corresponding to motor planning with coherence variables follow clear Gaussian-like distributions, while histograms corresponding to the motor planning process that involves categorical perception tend to decrease slower that the Gaussian distributions. Note that, even though histograms corresponding to the motor planning process that involves categorical perception are expected to be close to uniform in the 3-dimensional formant space, they are not so when projected, by marginalization, onto their dimensions, as is necessary for visualization purposes. Indeed, 3-dimensional uniform regions, when projected, become non-uniform

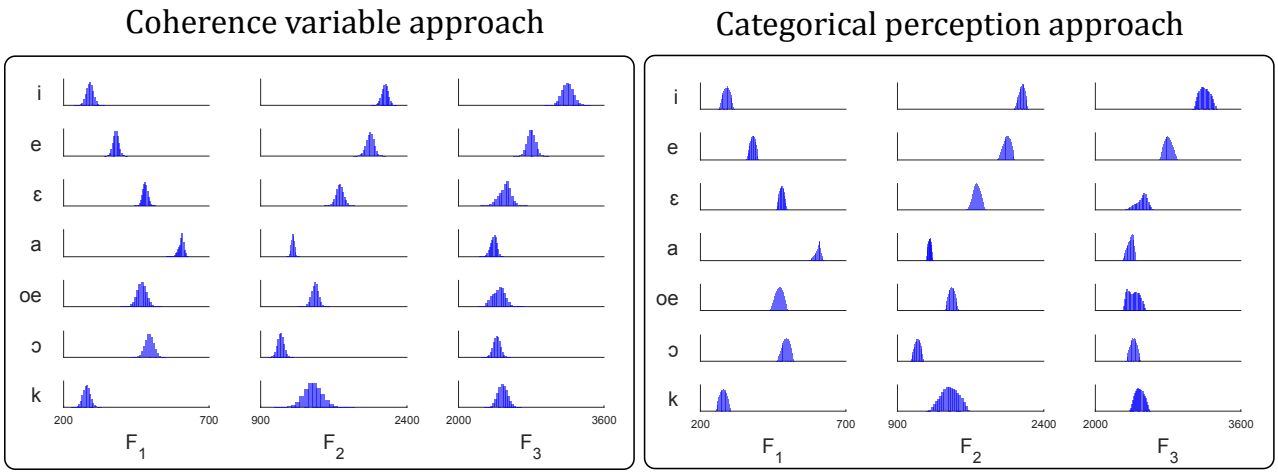

Figure 6.2: Comparison of auditory distributions obtained by the two model variants, involving or not categorical perception sub-module. *Histograms of auditory outputs resulting from samples M obtained from the motor planning processes involving coherence variables (left panels) and the motor planning processes involving categorical perception (right panels).*

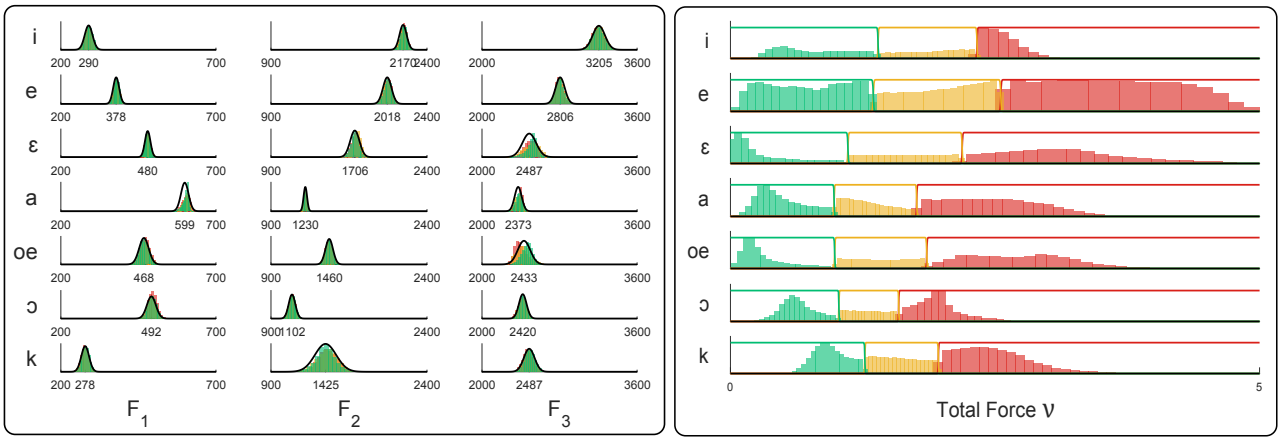

Figure 6.3: Histograms of auditory consequences (left panels) and forces (right panels) resulting from samples of M obtained from the motor planning processes $P(M \mid \Phi\ [C_A = 1]\ W)$ for all phonemes and effort levels.

because of the 3-d geometry of the region. Imagine for instance projecting a uniformly-dense sphere onto a plane: it would result in a non-uniform, bell-shaped curve over a circle, as the center would receive more "mass" than border regions.

3.2.2 Confirming the achievement of intended phonemes with intended levels of efforts

In Chapter 5 we showed that the motor planning process based on categorical perception was able to satisfy both the perceptual and effort constraints, for all phonemes and effort levels. Now we present the same evaluation for the planning process $P([M = m] \mid \Phi\ [C_A = 1]\ W)$, as defined in Eq (6.5).

Fig 6.3 presents histograms of auditory and force consequences of motor samples obtained from the motor planning process $P(M \mid \Phi\ [C_A = 1]\ W)$, for all phonemes and effort levels.

It can be seen that in all cases, the resulting auditory and effort samples correctly distribute according to the expected auditory force regions.

4 Discussion

We have defined two alternative models, based on the same pieces of assumed knowledge, but combined in different ways. The first approach leads to a motor planning process that is driven by categorical perception, which is mathematically defined as the result of an inference in a sub-model. The second approach leads to a planning process that is driven by the direct auditory specification of phonemes, which is mathematically defined as a coherence-constraint satisfaction between sensory-motor predictions and sensory expectations of phonemes.

Both models lead to qualitatively similar results, but differ however in the detailed distribution of motor outcomes that they predict. The first model predicts that productions of isolated phonemes distribute homogeneously inside the perceptual boundaries of the intended category, since categorical perception does not distinguish between central and border values. This results from the "inversion" of Gaussian prototypes (from $P(A \mid \Phi)$ to $P(\Phi \mid A)$), that yields target plateaus in the acoustic space. In contrast, the second model predicts a distribution of productions that matches the expected auditory distribution associated with each phoneme. Since we defined Gaussian distributions for these expectations, the model predicts a distribution of productions that is concentrated towards its center.

To our knowledge, no experimental study has specifically addressed this question, and it is unclear whether the variability of productions of isolated phonemes for a single speaker is rather uniform or Gaussian (for a study in the multi-speaker case, see Pijpers, Alder, and Togneri (1993)). We have designed and run an experimental study in order to evaluate this question [2]. However, preliminary results appear inconclusive, and this experiment warrants additional work; we therefore do not discuss it further here.

At this point we are currently unable to select one particular model over the other based on experimental observations. However, we choose to retain the approach involving the coherence variable for two theoretical reasons. Firstly, it is more flexible and general, as it subsumes the one involving sub-modules. Indeed, it is also possible to define a third variant, using both coherence variables and categorical planning, by introducing the coherence constraints and defining targets by inversion in the sub-module. Secondly, the coherence variable approach enables to connect or disconnect the related variables, which allows controlling explicitly the transfer of information in the model. This is a particularly useful property that we will use in next chapters.

[2] We thank Alexis Favre-Felix and Margaux Sutre for conducting the experiment, and Saïd Zeggai for data processing and performing statistical analyses.

Part II

Multisensory characterization of speech motor goals

Chapter 7

Extension of GEPPETO

1 Introduction

In the previous chapters we have illustrated how the Bayesian approach enables to formalize the intrinsic variability of speech at the level of motor planning, and how it enables to implement the same hypothesis as GEPPETO in order to account for two aspects of contextual variability: coarticulation and speech rate effects. We have proceeded in a step-by-step manner, by introducing first a single-phoneme planning model integrating perceptual and effort constraints, and then by extending it towards the planning of sequences of phonemes. This has illustrated the ease with which our modeling framework accommodates modular and hierarchical representation of knowledge.

Here, we present our first extension of B-GEPPETO that was not previously featured as is in GEPPETO, by considering both auditory and somatosensory information during speech planning. Indeed, compensatory behaviors reported by studies of auditory and somatosensory perturbations in speech indicate that both types of sensory information are taken into account during speech production (Houde & Jordan, 1998; Tremblay et al., 2003). However, these studies also report inter-individual differences in the amount of compensation to these perturbations, some subjects compensating more than others, some subjects not compensating at all, and almost no subject fully compensating.

The current version of GEPPETO takes into account both auditory and somatosensory information, however it does so at two different levels. Somatosensory information is taken into account at the low level of the muscle stretch reflex (see Chapter 4), which enables GEPPETO to account for online compensatory behavior to transient force perturbations (Szabados, 2017).

In the case of expected perturbations, subjects not only compensate by online error correction, but also adapt to the perturbation by learning new control patterns. Such adaptation to somatosensory perturbations cannot easily be accounted for in GEPPETO, since somatosensory feedback is only low-level, and adaptation at this level cannot be straightforwardly taken into account at the planning stage.

We therefore propose an extension of the B-GEPPETO model in order to consider multimodal representations of goals, so that auditory and somatosensory information are represented at equivalent levels. This chapter is adapted from Patri, Diard, and Perrier (2016). Section 2 begins by describing how we define the somatosensory space in the model. We then present, in Section 3, the multisensory extension of GEPPETO that involves motor goal regions in both auditory and somatosensory spaces. Section 4 presents simulation results provided by this multisensory model.

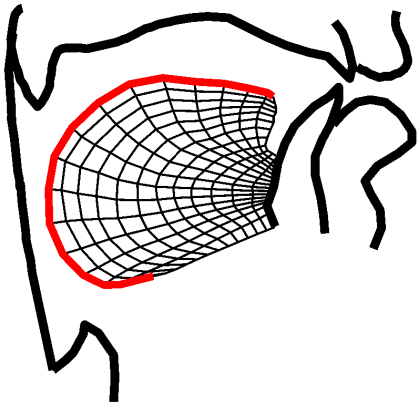

Figure 7.1: Specification of tongue contours in terms of the positions of the 17 surface nodes of the biomechanical model of the tongue (connected by the red curve). *Since each node is in 2-dimensional sagittal plane, each tongue configuration corresponds to a point in a 34-dimensional space.*

2 Defining somatosensory space

2.1 Context

The characterization of speech targets in somatosensory terms raises a certain number of theoretical and practical questions. We have defined auditory space in terms of the 3-dimensional formant space. How should we define the somatosensory space?

Somatosensory information can be divided into tactile and proprioceptive information. Tactile information is provided by cutaneous mechanoreceptors, and is present whenever there is contact, pressure or vibration of oral tissues. This source of information is certainly present for most consonants, where for instance stop or fricatives correspond to contact or vibration of the lips, tongue or palate. However, in the case of most vowels, tactile information is poor or absent and tongue configuration can only be assessed via proprioception.

Limb proprioception is generally assumed to be provided by sensory receptors in muscles (muscle spindles), tendons (golgi tendon organ) and joints (joints receptors). However, for oral articulators, little is known about the precise nature of sensory receptors. Indeed, the jaw is the only oral articulator where known sensory receptors have been reported (Brunner et al., 2011; Feng, Gracco, & Max, 2011; M. Honda, Fujino, & Kaburagi, 2002). Yet, it is believed that other kinds of sensory receptors may also provide information concerning articulatory configurations related to speech, for instance cutaneous mechanoreceptors may inform about skin stretch during jaw and lip movements.

Since we mainly focus on vowels that do not involve lip or jaw movements and for which tongue contact are limited, we consider only tongue proprioception. Yet, we still need to define a mathematical space characterizing tongue proprioception, the equivalent of formant space in the case of auditory information. Since proprioception informs about tongue configurations, a first alternative is to identify proprioceptive space with the space of tongue contours, defined in terms of the 17 surface nodes of the biomechanical model of the tongue (Fig 7.1). Since each node is located in the 2-dimensional sagittal plane, the resulting proprioceptive space is embedded in a 34-dimensional space. However, as they are constrained by tongue structure,

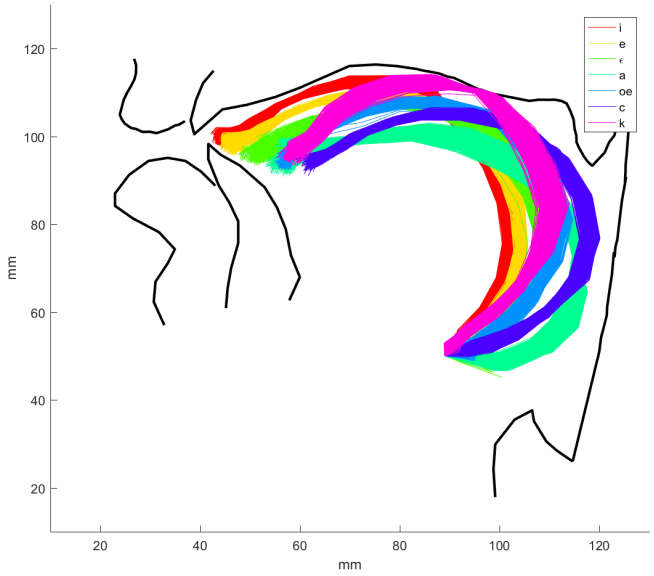

Figure 7.2: Tongue configurations retained for PCA analysis.

nodes are not independent from each other. Therefore, the actual space of tongue configurations is likely to be of much smaller dimension.

It is unclear however whether characterizing tongue proprioception in terms of tongue contours is biologically plausible, as little is known about the kind of sensory information that would enable to represent tongue contours. An alternative approach would be to consider that tongue proprioception may be provided by sensory receptors in tongue muscles that provide information about their length. Since we consider 6 groups of muscle fibers in the biomechanical model, we can therefore define proprioceptive space as this 6-dimensional muscle length space. However again, since muscles are constrained by tongue structure, the actual space of tongue configurations is likely to be of smaller dimension.

We have performed simulations of the model with both approaches, and it turns out that they lead to qualitatively similar results. The approach based on tongue contours lead to slightly better empirical results however. It is also more convenient for visualization, as tongue contours are easier to interpret than muscle lengths. For these reasons, we retain the approach based on tongue contour for the remaining of this work.

2.2 Assessing the dimensionality of somatosensory space

In order to assess the effective dimensionality of proprioceptive space, we performed a Principal Component Analysis (PCA) of the 34-dimensional description of tongue configurations. PCA was performed over 600 occurrences of each of the seven phonemes considered in the model. The data we used was extracted from the dictionary used to train the RBF networks of GEPPETO. This dictionary contains 27,615 instances of λ control variables, along with the corresponding auditory outputs, generated forces, tongue contours and muscle lengths. Among these 27,615 samples, around 7,000 correspond to the 7 auditory regions characterizing phonemes in the model. 600 is the smallest number of samples contained in one category. We randomly selected 600 samples among the other categories in order to take the same number of samples in each phoneme category.

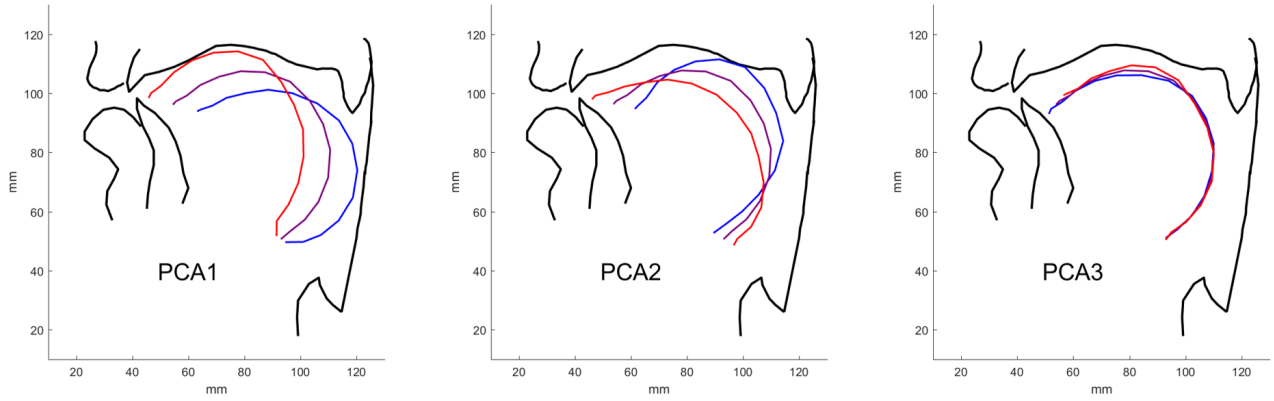

Figure 7.3: Effect of each of the three retained PCA components on tongue shapes.

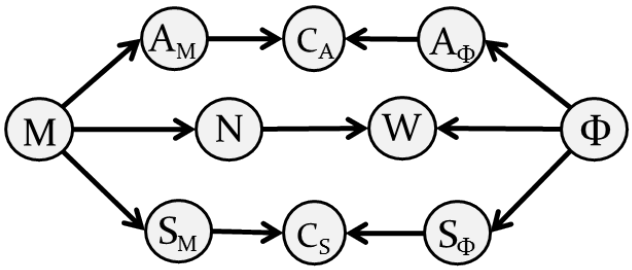

Figure 7.4: Graphical representation of the model with multisensory motor goals. *The diagram represents the decomposition of the joint probability distribution given in Eq (7.1).*

Fig 7.2 illustrates the tongue configurations used for the PCA analysis. The first three principal components describe more than 95% of the variance (74.1% for the first dimension, 92.3% for the first two dimensions, 96.1% for the first three dimensions). Hence, instead of describing tongue contours as points in a 34-dimensional space, we keep the three first PCA components and represent tongue configurations as points in this 3-dimensional PCA space. The effect of each PCA component on tongue shapes is illustrated in Fig 7.3. It can be seen that the first component corresponds to front-back configurations, the second component to low-high configurations and the third component to small vertical displacements of tongue body.

3 Bayesian model with auditory and somatosensory targets

3.1 Model definition

3.1.1 Variables

We build up on the previous models and therefore consider the same set of variables presented in Chapters 5 and 6. Total muscle force and effort levels are represented by variables N and W; control variables and their auditory predictions are represented by variables M and A_M; phonemes and their auditory expectations are represented by variables Φ and A_Φ; a coherence variable C_A enables to "connect" the two auditory variables A_M and A_Φ. In order to include somatosensory information, we further introduce variables S_M and S_Φ, representing respec-

tively the predicted somatosensory consequences of tongue configurations associated with motor commands M, and the expected somatosensory values characterizing phonemes Φ. S_M and S_Φ are both 3-dimensional continuous vector variables corresponding to tongue configurations described in the three dimensional PCA space defined in Section 2. As for auditory variables, we define an additional coherence variable C_S in order to "connect" the two somatosensory variables S_M and S_Φ. Similar to the auditory coherence variable C_A, the somatosensory coherence variable C_S acts as a binary switch (taking values {0,1}).

3.1.2 Decomposition

We decompose the joint probability distribution in the following way:

$$\begin{aligned} & P(M\ A_M\ A_\Phi\ C_A\ \Phi\ S_M\ S_\Phi\ C_S\ N\ W) \\ = \ & P(M)P(\Phi)P(A_M \mid M)P(A_\Phi \mid \Phi)P(C_A \mid A_M\ A_\Phi) \\ & P(N \mid M)P(W \mid \Phi\ N) \\ & P(S_M \mid M)P(S_\Phi \mid \Phi)P(C_S \mid S_M\ S_\Phi). \end{aligned} \tag{7.1}$$

Terms on the first two lines correspond to the decomposition given in Chapter 6 for the purely auditory model involving coherence variables; terms on the third line introduce the additional somatosensory variables. Fig 7.4 illustrates the decomposition given by Eq (7.1). Motor commands M and phonemes Φ appear as two pivotal nodes relating two otherwise independent sensory pathways, along with the pathway concerning forces and effort levels.

3.1.3 Parametric forms

We have already defined all terms of the first line of the decomposition in Eq (7.1). The three remaining terms correspond to the additional somatosensory knowledge that we are including in the model. These terms are defined as follows:

$\boldsymbol{P(S_M \mid M)}$ corresponds to the knowledge relating control variable M to its predicted somatosensory consequence S_M. This term is the somatosensory equivalent of the auditory-motor internal model $P(A_M \mid M)$, and therefore we identify it to a somatosensory-motor internal model characterized by Dirac distributions centered on a somatosensory-motor mapping ρ_s:

$$P([S_M = s] \mid [M = m]) \ \coloneqq \ \delta(s - \rho_s(m)) \ , \tag{7.2}$$

where δ denotes the Dirac distribution, translating the fact that $P([S_M = s] \mid [M = m])$ is zero unless $s = \rho_s(m)$. As for as ρ_a and ρ_ν in Chapter 5, ρ_s is implemented by a Radial Basis Functions network and is learned from simulations obtained with the biomechanical model of the tongue.

$\boldsymbol{P(C_S \mid S_M\ S_\Phi)}$ implements the sensory matching constraint, similar to $P(C_A \mid A_M\ A_\Phi)$, relating the two somatosensory variables S_M and S_Φ:

$$P([C_S = 1] \mid [S_M = s_m]\ [S_\Phi = s_\phi]) \ \coloneqq \ \begin{cases} 1 & \text{if } s_m = s_\phi \\ 0 & \text{otherwise.} \end{cases} \tag{7.3}$$

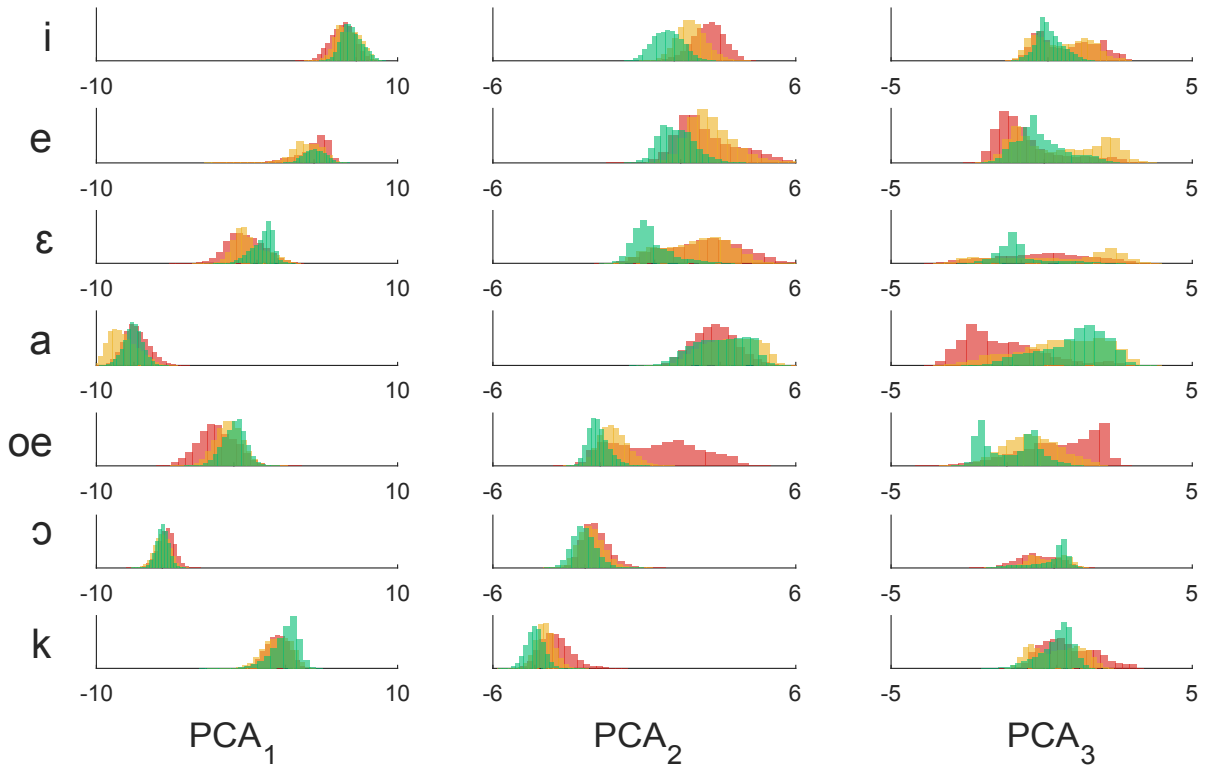

Figure 7.5: Histograms of somatosensory correlates of samples M obtained from the auditory planning $P(M \mid \Phi\ W\ [C_A = 1])$ for all phonemes and effort levels. *Green histograms correspond to effort level* W *="Weak", yellow histograms correspond to effort level* W *="Medium" and red histograms correspond to effort level* W *="Strong". Black curves represent the retained Gaussian distributions identified to the somatosensory characterizations of phonemes.*

$\boldsymbol{P(S_\Phi \mid \Phi)}$ corresponds to the knowledge relating phonemes to their expected somatosensory correlates. It is the somatosensory-equivalent of the auditory characterization of phonemes, identified with $P(A_\Phi \mid \Phi)$ in Chapter 6, which we defined as Gaussian probability distributions characterized by the auditory ellipsoid regions of GEPPETO. In the present case, we assume that the specification of $P(S_\Phi \mid \Phi)$ is learned from production driven by these auditory targets alone. More precisely, we assume that auditory characterization of phonemes, $P(A_\Phi \mid \Phi)$, are learned first and guide the initial productions through motor planning defined as in the previous model $P(M \mid \Phi\ W\ [C_A = 1])$. Progressively, the somatosensory correlates associated with these first productions are used to build the somatosensory characterization of phonemes $P(S_\Phi \mid \Phi)$.

We implement this learning mechanism in the model by identifying $P(S_\Phi \mid \Phi)$ with the inference $P(S_M \mid \Phi\ [C_A = 1]\ W)$ in the model. Performing this inference gives:

$$P(S_M \mid \Phi\ [C_A = 1]\ W) \quad \propto \quad \sum_M P(M \mid \Phi\ [C_A = 1]\ W) P(S_M \mid M). \tag{7.4}$$

The distributions given by Eq (7.4) cannot be computed exactly. However, samples can be obtained following an ancestral sampling approach (Bishop, 2006). Indeed, we can begin by sampling values of variable M according to $P(M \mid \Phi\ [C_A = 1]\ W)$ with the motor planning process defined in Chapter 6. Next, for each obtained motor value m we

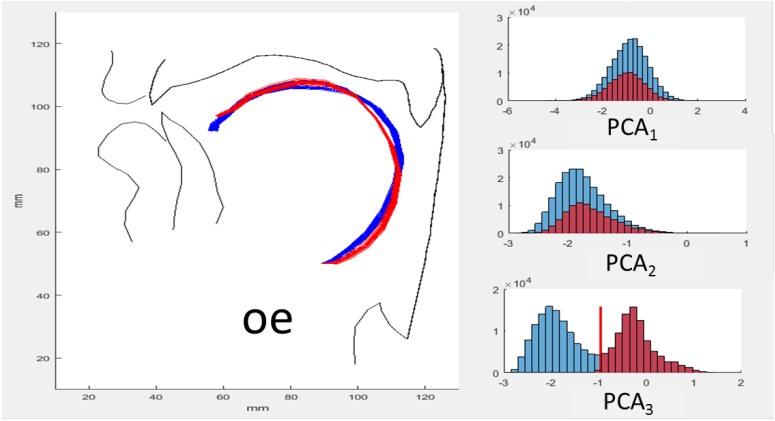

Figure 7.6: Two close but different tongue configurations leading to phoneme /œ/. *Right panels: histograms of somatosensory samples obtained for phoneme /œ/ with effort level W = "Weak" projected in the three PCA components (same histograms as the fifth line in Fig 7.5); samples are sorted according to the two modes in the third PCA component. Left panel: tongue contours corresponding to the sorted samples.*

can derive its somatosensory correlate s via the somatosensory-motor mapping ρ_s, since $P(S_M \mid M)$ are Dirac distributions centered on values predicted by ρ_s. The resulting somatosensory samples s then distribute according to $P(S_M \mid \Phi\ [C_A = 1]\ W)$.

Fig 7.5 presents histograms of somatosensory samples obtained as described, for all phonemes and effort levels. It can be seen that the resulting samples distribute slightly differently for different levels of effort. In particular, the effort level W="Strong" leads to distributions of samples that depart more strongly from the two others (e.g. for phonemes /i/, /a/ and /œ/). However, it is interesting to note that in most cases the resulting samples roughly follow Gaussian-like distributions, the case of effort level $W =$ "Weak" being the most clear. Yet, a notable exceptions can be seen in phoneme /œ/ where the projection into the third PCA component is a bimodal distribution. This reflects that, in this particular case, there are two close but different tongue configuration for the realization of this phoneme, as illustrated by Fig 7.6.

These results suggest that the somatosensory characterization of phonemes, $P(S_\Phi \mid \Phi)$, can be identified with Gaussian approximations to the somatosensory distributions obtained by auditory planning, $P(S_M \mid \Phi\ [C_A = 1]\ W)$. Since the effort level $W =$ "Weak" is the one that leads to more Gaussian-like distributions, we choose to identify $P(S_\Phi \mid \Phi)$ to Gaussian approximations of samples obtained with this effort level, i.e.:

$$P(S_\Phi \mid \Phi) = P(S_M \mid \Phi\ [C_A = 1]\ [W = \text{"Weak"}]). \tag{7.5}$$

The ellispoid regions on the right panel of Fig 7.7 indicate the position and geometry of the retained Gaussian distributions for each phoneme, in the two first principal components subspace of tongue shapes. The left panel of Fig 7.7 represent the auditory target regions for reference. Data points within increasing elliptical rings in auditory space have identical colors in the auditory and somatosensory spaces, providing an intuitive idea of the geometry distortion from the auditory to the somatosensory space.

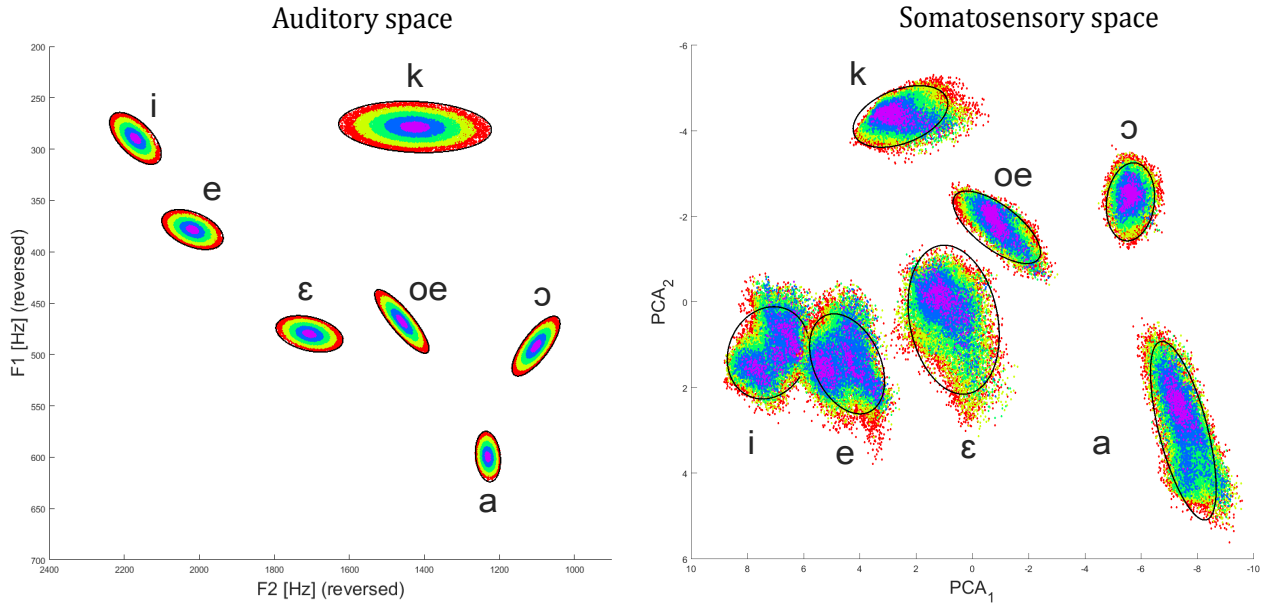

Figure 7.7: Auditory and somatosensory target regions. *Left Panel: ellipses representing auditory targets in* (F_2, F_1) *sub-space. Right panel: image of auditory target regions in somatosensory space; ellipses represent the Gaussian distributions (2 standard deviations) characterizing somatosensory targets. Colors enable to visualize the distortion of geometry between auditory and somatosensory spaces.*

3.2 Question: Inference of values of the control variables M

Coherence variables C_A and C_S allow to connect or disconnect the corresponding auditory and somatosensory pathways. This enables to define three different motor planning processes, depending on whether C_A, C_S, or both, are activated with value 1:

1. Activating only variable C_A leads to:

$$\begin{aligned} &P([M = m] \mid \Phi\ W\ [C_A = 1]) \\ &\quad \propto \quad P([A_\Phi = \rho_a(m)] \mid \Phi) P(W \mid \Phi\ [N = \rho_\nu(m)]). \end{aligned} \tag{7.6}$$

   This result depends only on knowledge involved in the auditory pathway. Hence, it correspond to a purely auditory planning mode (note that this result is the same as obtained in our previous purely auditory model).

2. Activating only variable C_S leads to:

$$\begin{aligned} &P([M = m] \mid \Phi\ W\ [C_S = 1]) \\ &\quad \propto \quad P([S_\Phi = \rho_s(m)] \mid \Phi) P(W \mid \Phi\ [N = \rho_\nu(m)]), \end{aligned} \tag{7.7}$$

   which depends only on knowledge involved in the somatosensory pathway. Hence, it correspond to a purely somatosensory planning mode.

3. Activating both variables C_A and C_S leads to:

$$\begin{aligned} &P([M = m] \mid \Phi\ [C_A = 1]\ [C_S = 1]) \\ &\quad \propto \quad P([A_\Phi = \rho_a(m)] \mid \Phi) P([S_\Phi = \rho_s(m)] \mid \Phi) P(W \mid \Phi\ [N = \rho_\nu(m)]), \end{aligned} \tag{7.8}$$

which depends on knowledge involved in both sensory pathways. It is therefore a planning process that combines auditory and somatosensory targets by performing a fusion that takes the form of a multiplication of probability distributions.

Eqs (7.6), (7.7) and (7.8) are obtained by the application of Bayesian inference to the joint probability distribution given by Eq (7.1).

4 Simulation results

We have defined three different planning processes based on either the auditory pathway alone, the somatosensory pathway alone, or the fusion of both sensory pathways. We have already presented results corresponding to the auditory pathway in Chapter 6. In this section we evaluate the outcome of thee two other planning process. We begin by presenting results corresponding to the somatosensory planning process alone in Section 4.1. We first confirm that somatosensory planning correctly satisfies both somatosensory and effort constraints. Then we further evaluate the auditory consequences resulting from this purely somatosensory planning process, in order to assess whether it leads to similar results as the purely auditory planning. Section 4.2 then presents results concerning the planning process based on the fusion of sensory pathways and compare its outcome to the results obtained with the purely auditory and the purely somatosensory planning processes.

4.1 Somatosensory planning $P(M \mid \Phi\ W\ [C_S = 1])$

4.1.1 Satisfying somatosensory and effort constraints

The first question to evaluate is whether the planning process correctly satisfies the imposed somatosensory and effort constraints. Fig 7.8 illustrates somatosensory and forces histograms resulting from samples M obtained with the somatosensory planning process $P(M \mid \Phi\ W\ [C_S = 1])$, for all phonemes and effort levels. It can be seen that the resulting samples correctly distribute according to the intended somatosensory targets and effort regions.

4.1.2 Auditory consequences

We now verify whether the purely somatosensory planning process correctly generates auditory consequences that agree with the auditory characterization of phonemes. Fig 7.9 illustrates histograms of auditory consequences resulting from the samples M obtained with the somatosensory planning process illustrated in Fig 7.8, for all phonemes and effort levels. It can be seen that, in most cases, the resulting samples correctly distribute according to the intended auditory targets, indicated by the superposed black curves. Most errors are associated with "Strong" and "Medium" effort levels, in particular for F_3 values of phoneme /i/ /e/ and /k/. These differences can be surprising given the fact that the corresponding somatosensory samples distributed similarly for all levels of effort (Fig 7.8). The observed differences may originate from differences in greater PCA dimensions, or due to inaccuracies of the internal models in regions with strong non-linearities between auditory and somatosensory spaces. Further work would be needed to clarify these questions.

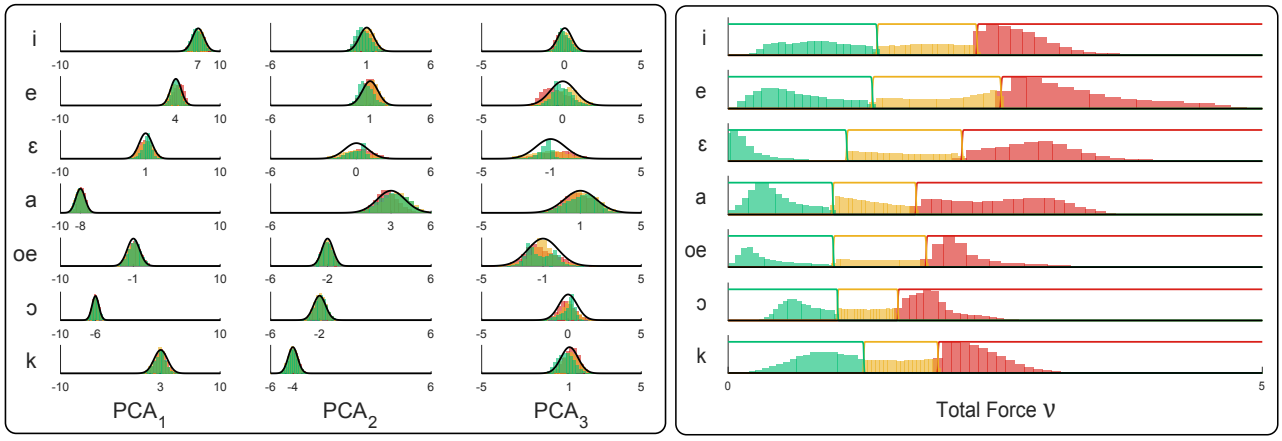

Figure 7.8: Histograms of somatosensory consequences (left panels) and forces (right panels) resulting from samples M obtained with the somatosensory planning process, $P(M \mid \Phi\, W\, [C_S = 1])$, for all phonemes and effort levels. *Green histograms correspond to effort level* W *="Weak", yellow histograms correspond to effort level* W *="Medium" and red histograms correspond to effort level* W *="Strong". Black curves on left panels represent the Gaussian distributions corresponding to somatosensory characterizations of phonemes; colored curves on the right panels indicate the corresponding regions characterizing each level of effort in GEPPETO.*

4.2 Sensory fusion planning $P(M \mid \Phi\ W\ [C_A = 1]\ [C_S = 1])$

The fusion planning process imposes all three constraints at the same time: auditory goals, somatosensory goals and intended effort levels. We begin by evaluating whether the inferred values of control variable M correctly satisfy both auditory and somatosensory goals, and then evaluate intended levels of effort.

4.2.1 Achieving both auditory and somatosensory goals

Fig 7.10 illustrates histograms of somatosensory and auditory correlates resulting from samples of M obtained with the fusion planning process $P(M \mid \Phi\, W\, [C_A = 1]\, [C_S = 1])$, for all phonemes and effort levels. It can be seen that all somatosensory and auditory samples correctly distribute according to the intended somatosensory and auditory motor goals. Note however that auditory and somatosensory histograms are slightly narrower compared to the intended distributions (indicated by black Gaussian curves). This is a consequence of the fact that the probabilistic fusion of two Gaussian distributions is a Gaussian distribution of smaller variance (Ernst & Banks, 2002). Even if our samples do not exactly follow Gaussian probability distributions, they are unimodal and well-localized in their respective spaces, so that the property of variance diminution by fusion also appears to hold.

4.2.2 Satisfying the effort constraint

Fig 7.11 illustrates histograms of samples of the total level of force N resulting from samples of M obtained with the fusion planning process $P(M \mid \Phi\, W\, [C_A = 1]\, [C_S = 1])$, for all phonemes and effort levels. As in all previous cases, it can be seen that the intended ranges of forces are satisfied for all phonemes and effort levels.

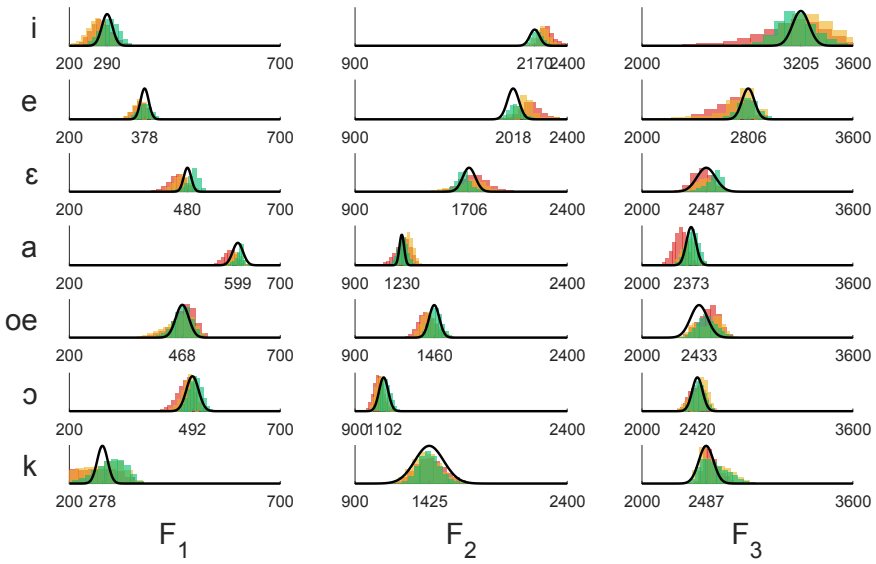

Figure 7.9: Histograms of auditory consequences of samples M obtained with the somatosensory planning process, $P(M \mid \Phi\ W\ [C_S = 1])$, for all phonemes and effort levels. *Green histograms correspond to effort level* $W =$ Weak*, yellow histograms correspond to effort level* $W =$ Medium *and red histograms correspond to effort level* $W =$ Strong*. Black curves represent the Gaussian distributions corresponding to auditory characterizations of phonemes.*

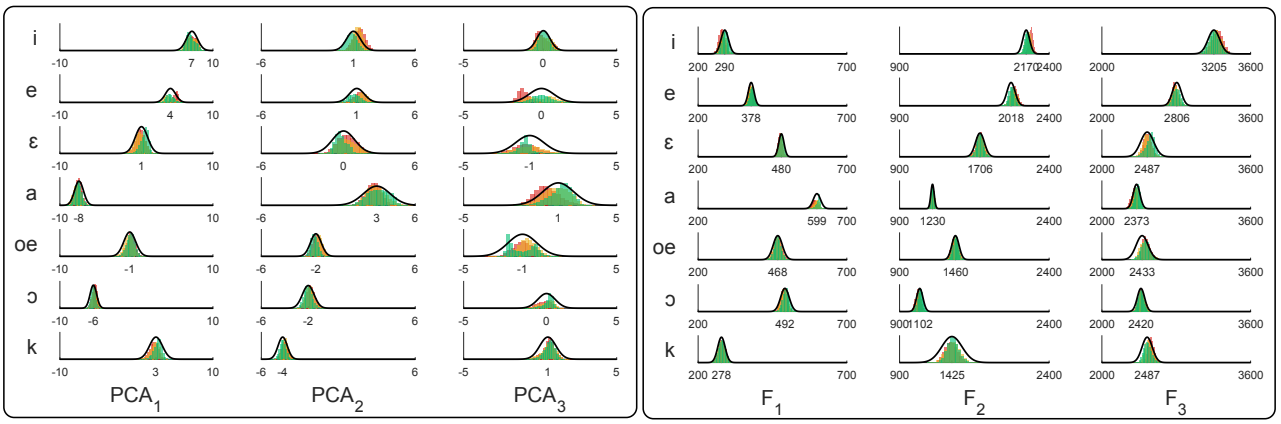

Figure 7.10: Histograms of somatosensory consequences (left panels) and auditory consequences (right panels) resulting from samples of M obtained with the fusion planning process, $P(M \mid \Phi\ W\ [C_A = 1]\ [C_S = 1])$, for all phonemes and effort levels. *Green* $W =$ Weak*; yellow* $W =$ Medium*; red* $W =$ Strong*. Black curves represent the Gaussian distributions corresponding to somatosensory and auditory characterizations of phonemes.*

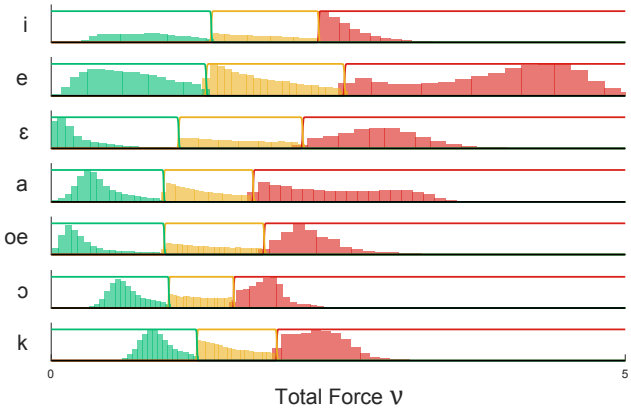

Figure 7.11: Histograms of samples of the total level of force N corresponding to samples of M obtained with the fusion planning process, $P(M \mid \Phi\ W\ [C_A = 1]\ [C_S = 1])$, for all phonemes and effort levels. *Green* $W = \text{Weak}$*; yellow* $W = \text{Medium}$*; red* $W = \text{Strong}$. *Superimposed colored curves specify the corresponding regions characterizing each level of effort in GEPPETO.*

Chapter 8

Modeling sensory preferences in speech motor planning

1 Introduction

Lametti et al. (2012) explored the idea that differences in the amount of compensation to sensory perturbations in speech production would be due to individual differences in how subjects integrate each sensory modality, some subjects relying more on auditory information, others on somatosensory information. In agreement with this hypothesis, Lametti et al. (2012) observed a negative correlation between the amount of compensation to perturbations in each sensory modality. In other words, subjects who compensated more to one sensory perturbation tended to compensate less to the other, suggesting the existence of individual sensory preferences in speech motor control.

Differences in the amount of compensation have been previously interpreted, in the context of the DIVA model, as differences in the size of sensory targets regions (J. S. Perkell et al., 2008; Villacorta et al., 2007). Under this interpretation, subjects with lower sensory acuity in one modality learn wider sensory goal regions and become less sensitive to sensory errors in that sensory modality. This interpretation implies that sensory preferences must be stable and could only change slowly if sensory goal regions change. However, sensory reliance may well not be a fixed property of each individual, but could rather be modulated by context, and hence could change rapidly. For instance, larger auditory perturbations appear to be less compensated than smaller perturbations (Katseff, Houde, & Johnson, 2012), and this may suggest that sensory reliance may be down-weighted in the case of large sensory errors (but see Hahnloser and Narula (2017); Sober and Brainard (2012); Wei and Kording (2009) for other interpretations in the context of arm reaching or bird songs).

In this chapter we explore how sensory preferences can be represented in the model, without necessarily relying on modifications of sensory goal regions. In order to address this question, we begin, in Section 2, by introducing how we implement sensory perturbations and adaptation in the model. Then, in Section 3, we evaluate two alternative ways of implementing sensory preferences in the model: in the first one, precision of each sensory target region is modulated, and in the second one, target regions are not affected but instead, the coherence constraint on each sensory modality is modulated. This suggests a theoretical extension of coherence variables, towards the continuous case, which we explore in Section 4.

2 Implementing sensory perturbations and adaptation in the model

2.1 General principle

Adaptation is interpreted in our framework as the update of parameters characterizing pieces of knowledge included in the model, that is to say, the terms of the decomposition of the joint probability distribution. These are the priors $P(M)$ and $P(\Phi)$, the two sensory-motor internal models, $P(A_M \mid M)$ and $P(S_M \mid M)$, and the two sensory characterizations of phonemes, $P(A_\Phi \mid \Phi)$ and $P(S_\Phi \mid \Phi)$. Sensory perturbations modify the sensed consequences of motor actions such that the predicted sensory consequences of motor commands become erroneous. When the perturbation is consistently maintained, a new relation between motor commands and corresponding sensory consequences is experienced. Therefore, we interpret adaptation as the update of the corresponding sensory-motor internal model, $P(A_M \mid M)$ or $P(S_M \mid M)$, in order to capture the new sensory-motor relation imposed by the perturbation.

We consider the case of adaptation to auditory and somatosensory perturbations, and analyze outcomes of the model only for the effort level $W =$ "*Weak*". However, as a short hand of notations, we will ignore the term [$W =$ "*Weak*"] in the expression of each planning process in this section. For instance, the auditory planning process for effort level [$W =$ "*Weak*"], $P(M \mid \Phi\ [W =$ "*Weak*"$]\ [C_A = 1])$, will be simply written $P(M \mid \Phi\ [C_A = 1])$.

2.2 Adaptation to an auditory perturbation

Two main sort of auditory perturbations have been considered in speech adaptation studies: delays of the auditory feedback (Stuart, Kalinowski, Rastatter, & Lynch, 2002; Yamamoto & Kawabata, 2014; Yates, 1965), and shifts of particular spectral characteristics of speech sounds, among which, shifts of subjects' fundamental frequency (pitch shift) (Burnett, Freedland, & Larson, 1998; Jones & Munhall, 2000), shifts of the first spectral moment for fricatives (Shiller et al., 2009), or shifts of one or more formants for vowels (Cai, Ghosh, Guenther, & Perkell, 2010; Houde & Jordan, 1998; Purcell & Munhall, 2006). Since we are focusing on vowels we focus on this last type of perturbation. More specifically, in the following, we consider a consistent shift of the first formant by -100 Hz, during the production of vowel /ɔ/. The middle panel of Fig 8.1 illustrates the effect of this perturbation on productions resulting from our three planning processes.

2.2.1 Implementing adaptation

We implement adaptation by updating the auditory motor mapping in normal condition, ρ_a^n, in order to take into account the perturbation. Since the perturbation simply shifts auditory values by a constant vector δ_A (in the present case $\delta_A = [-100, 0, 0]$), we define the updated internal model, ρ_a^u, as:

$$\rho_a^u(m) = \rho_a^n(m) + \delta_A. \tag{8.1}$$

2.2.2 Results

The right panel of Fig 8.1 illustrates the effect of the perturbation and the outcome of adaptation for each of the three planning processes. In normal conditions (left panels), all three planning correctly reach both auditory and somatosensory target regions. Since the perturbation is only

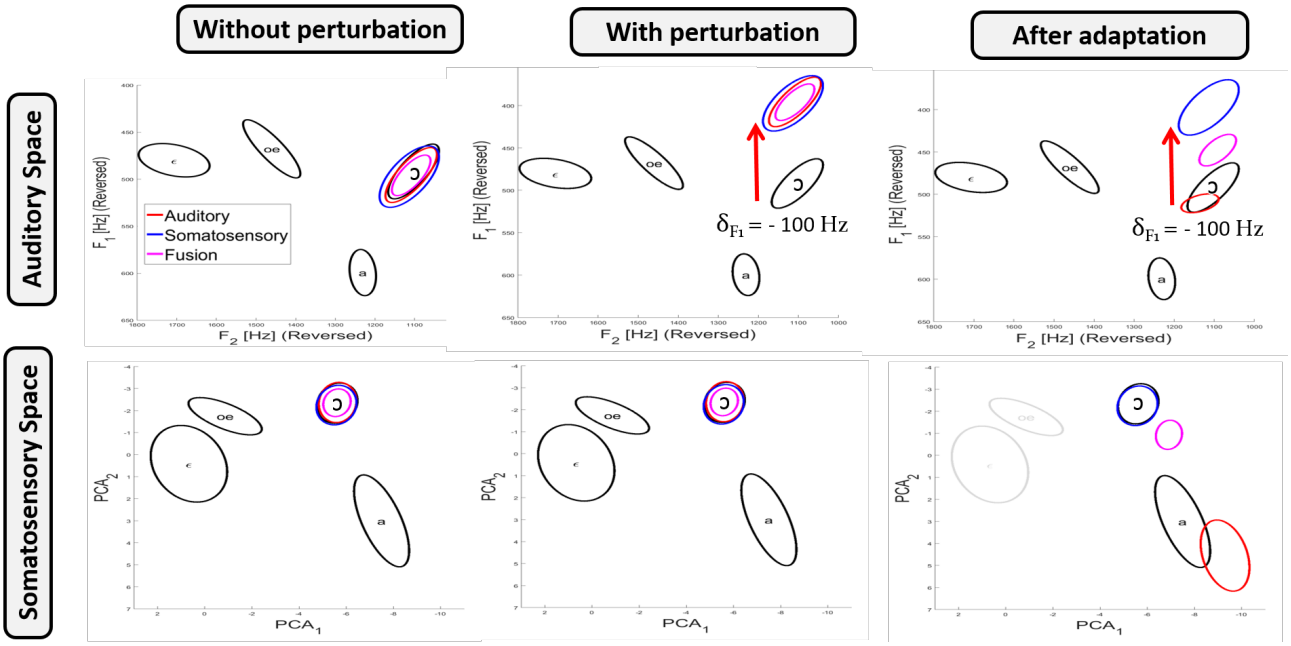

Figure 8.1: Effect of the auditory perturbation (middle panels) and result of adaptation (right panels) on the three planning processes – auditory (red), somatosensory (blue) and fusion (magenta) – for the production of phoneme /ɔ/, in auditory space (top panels) and somatosensory space (bottom panels).

auditory, it induces a shift of all planning processes only in auditory space (top middle panel) and not in somatosensory space (bottom middle panel).

Let us now evaluate the outcome of adaptation (right panels). Since somatosensory planning does not involve the auditory-motor mapping ρ_a (Eq 7.7), it is not affected by the update of the auditory-motor internal model and results remain unchanged. On the other hand, as expected, after updating the auditory-motor internal model (right panels), the auditory planning $P(M \mid \Phi\ [C_A = 1])$ fully compensates the perturbation and reaches again the auditory target region (top right panel). However, this compensation is performed by modifying articulatory patterns that therefore miss the corresponding target regions in somatosensory space (bottom right panel).

The fusion planning combines the two previous planning schemes. Since auditory and somatosensory target regions are now incompatible, fusion planning cannot reach both sensory targets at the same time, and therefore makes a compromise between them. As a result, fusion planning leads to auditory and somatosensory consequences that lie between those of auditory and somatosensory plannings.

2.3 Adaptation to a somatosensory perturbation

Auditory perturbations alter the relation between motor actions and expected auditory consequences. Similarly, somatosensory perturbations aim at altering the relation between motor actions and expected somatosensory consequences. This has been most commonly implemented with a force perturbation paradigm in the context of arm reaching movements, where a robotic device applies forces to the arm in order to alter its movements. In this case, the same control patterns lead to different articulatory results in the presence or absence of perturbation. In the context of speech, force perturbation paradigms have been mainly applied to jaw movements

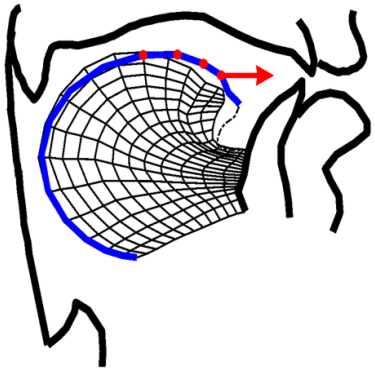

Figure 8.2: Force perturbation applied to the biomechanical model. *The red dots indicate the four nodes on which the force is applied. The mesh represents the tongue configuration without perturbation and the blue trace represents the contour of the tongue in the presence of the perturbation.*

(Lametti et al., 2012; Nasir & Ostry, 2009; Tremblay et al., 2003).

Since our model is based only on movements of the tongue, we consider a similar force perturbation paradigm, but applied to the tongue. The perturbation is implemented in the biomechanical model by including constant horizontal forces applied to four of the surface nodes of the tongue, as illustrated by Fig 8.2. These forces induce a displacement of the tongue that results from the interaction between the configuration of internal forces in the tongue and the direction and amplitude of the external perturbation.

2.3.1 Implementing adaptation

It is important to note that the present force perturbation is not totally symmetric to the previous auditory case. While the auditory perturbation altered only the auditory consequences of motor actions, the force perturbation on the tongue alters both somatosensory and auditory consequences. Indeed, the applied force alters the articulatory configuration of the tongue which leads to differences in both its auditory and somatosensory consequence. In order to implement a purely somatosensory perturbation and avoid the additional auditory effect, we assume a (hypothetic) force perturbation paradigm including an alteration of the auditory feedback in order to maintain the same auditory-motor relation as in normal conditions.

In this context, adaptation is implemented in the model by updating only the somatosensory motor mapping ρ_s characterizing the internal model $P(S_M \mid M)$. The new relation between motor commands and predicted somatosensory consequences is not as straightforward as in the previous auditory case however. As previously noted, the altered configuration of the tongue results from the interaction between forces generated by the tongue and the external perturbation. It results that the displacement of the tongue is not a simple constant shift along a single dimension, as in the auditory case. Therefore, we identify the updated somatosensory motor mapping ρ_s^u with a new RBF network trained on simulations of the biomechanical model with the force perturbation described previously.

2.3.2 Results

The right panel of Fig 8.3 illustrates the effect of the force perturbation and the outcome of adaptation for each of the three plannings for the production of phoneme /ɔ/. In normal

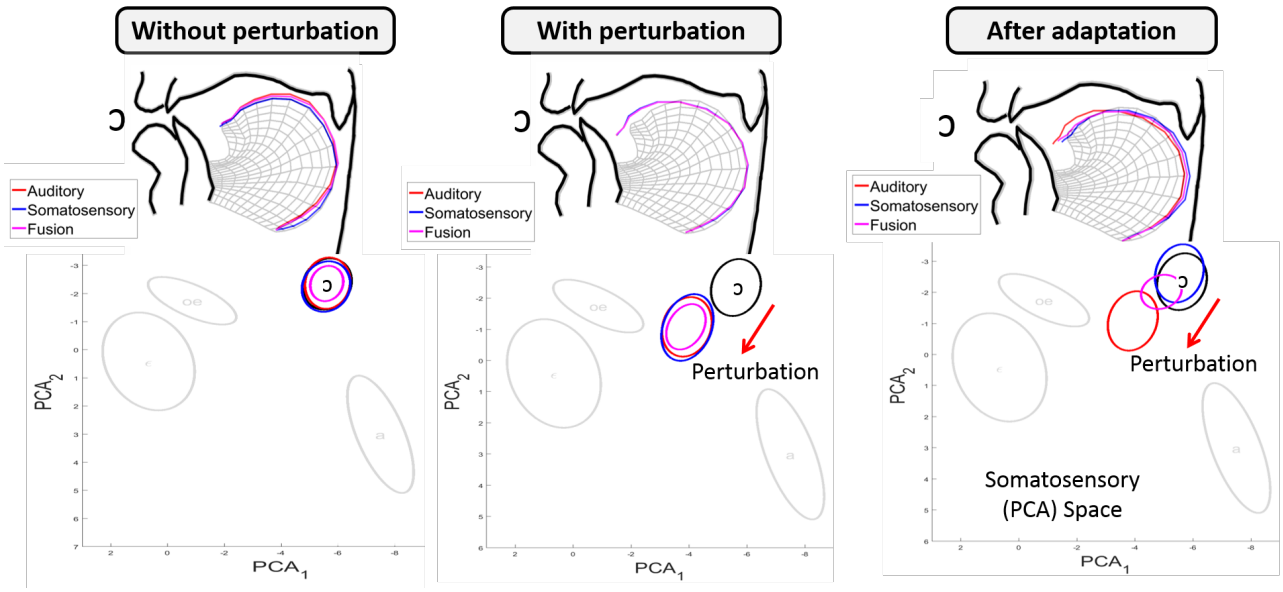

Figure 8.3: Effect of the force perturbation (middle panels) and result of adaptation (right panels) on the three planning processes – auditory (red), somatosensory (blue) and fusion (magenta) – for the production of phoneme /ɔ/. Top panels indicate the mean tongue postures obtained by each planning, the bottom panels represent the same results in somatosensory (PCA) space.

conditions (left panels), all three plannings correctly achieve both auditory and somatosensory target regions. When the force is applied (middle panels) the results obtained by all three plannings are displaced frontwards. Updating the somatosensory-motor mapping ρ_s (right panels) leads to similar but symmetric results as in the previous auditory case. This time it is the auditory planning $P(M \mid \Phi\ [C_A = 1])$ that is not affected by the update since it does not involve ρ_s (Eq 7.6). Also, it is the somatosensory planning $P(M \mid \Phi\ [C_S = 1])$ that fully compensates the perturbation and reaches again the somatosensory target region (top right panel) after updating the auditory-motor internal model (right panels). Finally, fusion planning combines again the two previous planning schemes and leads to auditory and somatosensory consequences that lie between those of the auditory and somatosensory planning schemes.

3 Implementing sensory preferences in the model: a study of two variants

We have implemented adaptation in the model and seen how the three planning schemes behave differently. The auditory and somatosensory planning schemes perform symmetrically, with one fully compensating the perturbation whereas the other does not compensate at all, and *vice versa*. The fusion planning scheme combines the auditory and somatosensory planning schemes, such that it always compensates the perturbation, but only partially. Our aim now is to study how the relative weights of auditory and somatosensory pathways could be modulated in the fusion planning in order to obtain a fusion planning scheme that would rely more on one or the other of the sensory pathways. In other words, our aim is to be able to represent sensory preference in the model.

In this section, we show that sensory preference can be implemented in two ways. In the

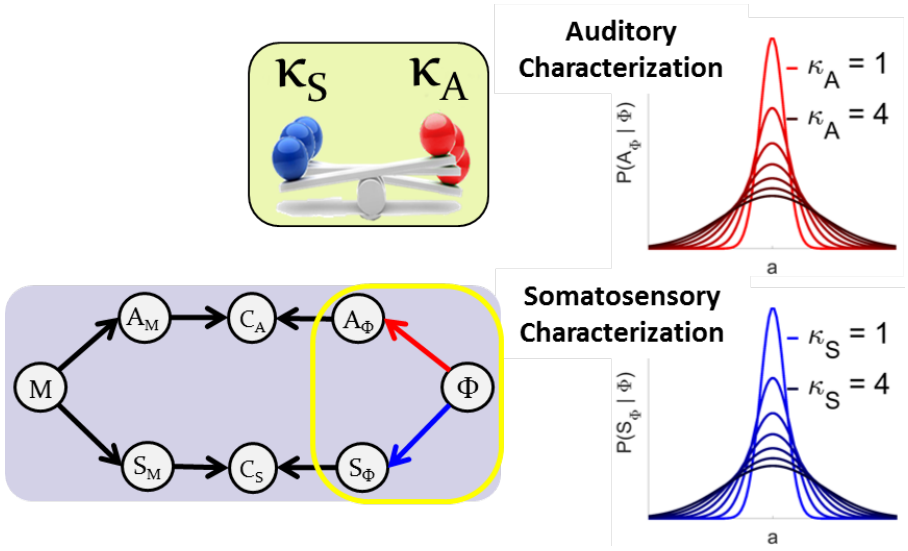

Figure 8.4: Illustration of the effect of parameters κ_A and κ_S on the auditory and somatosensory characterization of phonemes $P(A_\Phi \mid \Phi)$ and $P(S_\Phi \mid \Phi)$. *The greater the value of κ, the wider the distribution, and the weaker the contribution of the corresponding sensory pathway to the planning process.*

first variant, sensory preference is attributed to the relative precision of sensory regions characterizing speech motor goals, $P(A_\Phi \mid \Phi)$ and $P(S_\Phi \mid \Phi)$, with precision the inverse of covariance matrices Σ_A^ϕ and Σ_S^ϕ. This approach is inspired from classical models of multisensory fusion for perception (Ernst & Banks, 2002; J. S. Perkell et al., 2008) and follows the same idea as suggested in the DIVA model: precision of sensory regions correspond to their tolerance to perturbations; the smaller the region, the higher the precision and the lower the tolerance to perturbations. In other words, subjects who compensate more to auditory than somatosensory perturbations would have auditory target regions smaller than their somatosensory target regions.

In the second variant, sensory pathways are modulated by the sensitivity of the comparison between the predicted sensory consequences of the motor variable (A_M or S_M) and the sensory expectations associated with phonemes (A_Φ or S_Φ). This provides an alternative formulation of sensory preference that does not assume changes in the sensory characterization of phonemes.

3.1 First variant: modulate the precision of sensory targets

Implementation The precision of sensory targets can be modulated in the model by adjusting parameters κ_A and κ_S that multiply the covariance matrices of Gaussian distributions characterizing $P(A_\Phi \mid \Phi)$ and $P(S_\Phi \mid \Phi)$. For greater values of κ, the corresponding target region becomes wider, as illustrated in Fig 8.4.

Simulation results We evaluate the effect of parameters κ_A and κ_S in the context of adaptation to the auditory perturbation presented in Section 2. Fig 8.5 illustrates mean results obtained with motor planning under the fusion of sensory pathways, for different values of parameters κ_A and κ_S. Ellipses representing the results obtained in the previous section are represented for reference. Increasing parameter κ_A (resp. κ_S) leads to results that progressively drift towards the outcome of the somatosensory (resp. auditory) planning process.

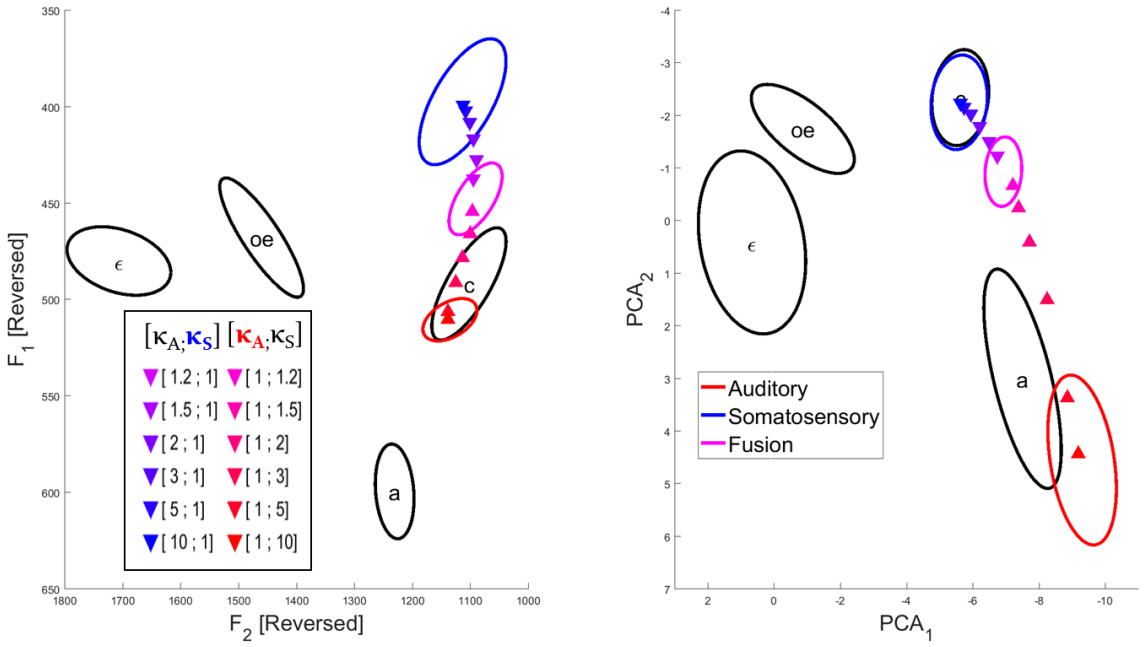

Figure 8.5: Results of the fusion planning process after adaptation to the auditory perturbation described in Section 2, for different values of parameters κ_A and κ_S.

Hence, parameters κ_A and κ_S effectively modulate the strength of each sensory pathways: the greater the value of κ, the smaller the contribution of the corresponding sensory pathway. This confirms the possibility of implementing sensory preferences in our model in a way similar to previous approaches: modulating the relative precision of sensory targets effectively modulates the contribution of the corresponding sensory pathway.

3.2 Second variant: modulate the sensitivity of the sensory matching constraints

Until now, we have defined the sensory matching constraints in an "all-or-nothing" manner:

$$P([C_S = 1] \mid [S_M = s_1]\ [S_\Phi = s_2]) \quad = \quad \begin{cases} 1 & \text{if } s_1 = s_2 \\ 0 & \text{otherwise,} \end{cases} \tag{8.2}$$

$$P([C_A = 1] \mid [A_M = a_1]\ [A_\Phi = a_2]) \quad = \quad \begin{cases} 1 & \text{if } a_1 = a_2 \\ 0 & \text{otherwise.} \end{cases} \tag{8.3}$$

These constraints are strict and rigid, in the sense that they remove all possible sensory values (giving them zero probability) unless they match exactly. Intuitively, if we are able to soften or relax this constraint, we may be able to modulate their strengths and hence control their involvement in the planning process.

Implementation Relaxing the constraint means that, instead of discarding all values unless they are exactly equal, we would allow values that are different but close, while still removing values that are indeed far from each other. In the case of the somatosensory-matching constraint (the auditory case being similar), this idea can be implemented by:

$$P([C_S = 1] \mid [S_M = s_1]\ [S_\Phi = s_2]) \quad = \quad e^{-d_S(s_1, s_2)}, \tag{8.4}$$

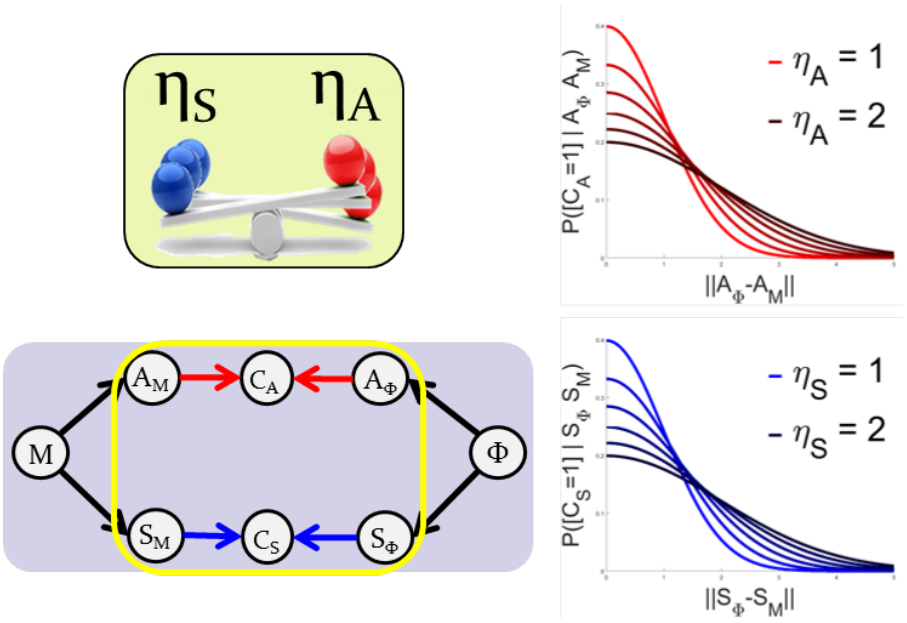

Figure 8.6: Illustration of the effect of parameters η_A and η_S on their corresponding sensory matching constraints. *The smaller the value of η, the sharper the constraint function and the stronger the relative contribution of the corresponding sensory pathway to the planning process.*

where $d_S(s_1,s_2)$ is a similarity measure, or distance, between values s_1 and s_2 in somatosensory space (Bessière et al., 2013). Since e^{-x} decreases continuously with increasing values of x, the function defined in Eq (8.4) gives high probability to values that are close (small distance $d_S(s_1,s_2)$) and low probability to values that are far from each other (large distance $d_S(s_1,s_2)$).

Eq (8.4) enables us to implement constraints of different strengths by specifying how sensitive the similarity measure d_S is for different sensory values. In the present case we choose a family of quadratic measures defined by:

$$d(a,b;\eta) \quad = \quad \frac{(a-b)^2}{2\eta^2}, \tag{8.5}$$

where η is a parameter that modulates the sensitivity of the measure.

With this choice of quadratic measure, Eqs (8.2) and (8.3) become:

$$P([C_S=1] \mid [S_M=s_1]\ [S_\Phi=s_2]) \quad = \quad e^{-\frac{(s_1-s_2)^2}{2\eta_S^2}} \tag{8.6}$$

$$P([C_A=1] \mid [A_M=a_1]\ [A_\Phi=a_2]) \quad = \quad e^{-\frac{(a_1-a_2)^2}{2\eta_A^2}}. \tag{8.7}$$

Fig 8.6 illustrates the form of the constraint functions defined by Eqs (8.6) and (8.7) for different values of parameter η: small (resp. large) values of η lead to sharper (resp. flatter) matching constraints. Note in particular that for $\eta \to 0$ the rigid constraints of Eqs (8.2) and (8.3) are recovered, while for $\eta \to +\infty$ the constraint functions become constant, independent of sensory values, and therefore correspond to an absence of constraint.

Simulation results We have formulated a generalization of the sensory matching constraints that enables to modulate their strength with sensitivity parameters η_S and η_A for somatosensory

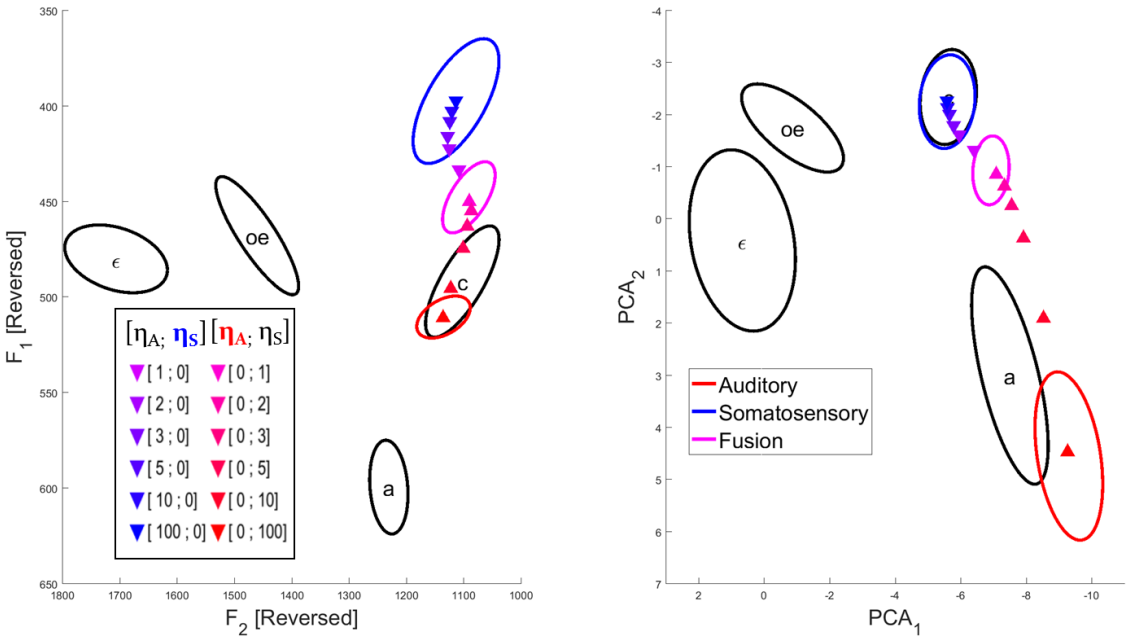

Figure 8.7: Results of the fusion planning process after adaptation to the auditory perturbation described in Section 2, for different values of parameters η_A and η_S.

and auditory pathways respectively, while including the hard constraints defined in Eqs (8.2) and (8.3) as special cases.

We evaluate the effect of parameters η_S and η_A in the context of adaptation to the auditory perturbation as in the previous approach. Fig 8.7 illustrates mean results obtained with motor planning under the fusion of sensory pathways, for different values of parameters η_S and η_A. Results obtained by the purely auditory and purely somatosensory plannings (as in Section 2) are represented for reference.

Increasing parameter η_A of the auditory matching constraint (resp. η_S of the somatosensory matching constraint) leads to results that progressively drift towards the outcome of the somatosensory (resp. auditory) planning process. Hence, parameters η_A and η_S also enable to modulate the strength of the constraint imposed by the corresponding sensory pathways, similarly to parameters κ_A and κ_S in the previous approach.

3.3 Discussion

Equivalence of approaches We have formulated two alternative approaches for modulating the relative weights of auditory and somatosensory pathways in the model. Simulations indicate that both approaches lead to very similar results and actually, this similarity can be proven formally. Indeed, computing the outcome of fusion planning with the generalized sensory matching constraints, given by Eqs (8.6) and (8.7), leads to:

$$P([M=m] \mid \Phi\ [C_A=1]\ [C_S=1]) \ \propto\ \mathcal{N}_S(\rho_s(m) \mid \Phi,\, \eta_S)\ \mathcal{N}_A(\rho_a(m) \mid \Phi,\, \eta_A), \qquad (8.8)$$

where $\mathcal{N}_S$ and $\mathcal{N}_A$ are Gaussian probability distributions defined as:

$$\mathcal{N}_S(x \mid \Phi,\, \eta_S) \ = \ \mathcal{N}(x\ ;\ \mu_S^\Phi,\, \Gamma_S^\Phi + \eta_S^2 I_{n_S}) \qquad (8.9)$$

$$\mathcal{N}_A(x \mid \Phi,\, \eta_A) \ = \ \mathcal{N}(x\ ;\ \mu_A^\Phi,\, \Gamma_A^\Phi + \eta_A^2 I_{n_A}). \qquad (8.10)$$

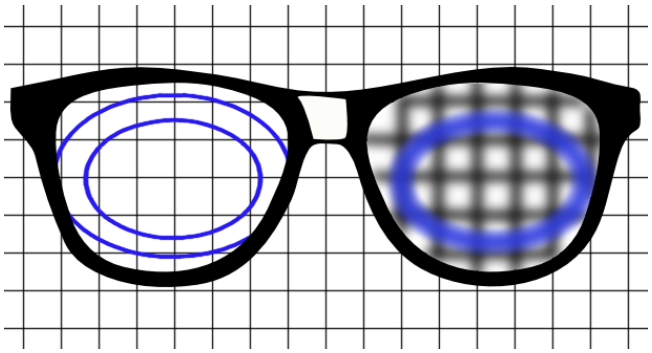

Figure 8.8: Illustrative interpretation of the formal equivalence between the two implementations of sensory preferences. *In the first approach (left part of the figure) auditory or somatosensory target regions are directly increased with parameters κ_A and κ_S. In the second approach (left part of the figure) parameters η_S and η_A modulate the sensitivity of the corresponding sensory matching constraint. Increasing parameters η reduce the sensitivities inducing a "blurring" effect on the corresponding sensory pathways which distort the target as if it was seen through misted glasses. As a consequence, the resulting sensory target appears wider than it actually is. Therefore, increasing the "blurriness" of the pathway through parameter η_S or η_A turns out to be equivalent to increasing the size of the corresponding sensory target regions with parameter κ_S or κ_A.*

Parameters (μ_A, Γ_A) and (μ_S, Γ_S) are the mean and covariance matrices of the sensory characterization of phonemes, $P(A_\Phi \mid \Phi)$ and $P(S_\Phi \mid \Phi)$. I_{n_A} and I_{n_S} are identity matrices of rank n_A and n_S (n_A and n_S being the number of dimensions of auditory and somatosensory spaces respectively).

Eq (8.8) is similar to the fusion planning scheme of Eq (7.8) in Chapter 7, with the strict coherence constraints. However, while the previous planning process was proportional to the product of two Gaussian distributions, corresponding to the two sensory characterization of phonemes, $P(S_\Phi \mid \Phi)$ and $P(A_\Phi \mid \Phi)$, this new planning process is proportional to the product of Gaussian distributions with same means, but covariance matrices increased by parameters η_A and η_S. Hence, parameters η_A and η_S play a similar role as parameters κ_A and κ_S in the previous approach: while parameters κ increased the covariance of the sensory characterization of phonemes multiplicatively, parameters η increase them additively. Fig 8.8 gives an intuitive interpretation of this result.

Finally, note that if auditory and somatosensory spaces were 1-dimensional, both approaches would be exactly equivalent, since any additive increase $\Gamma + \eta$ can be written as a multiplicative increase $\kappa\Gamma$, with $\kappa = 1 + \frac{\eta}{\Gamma}$. This is not true anymore in higher dimensions though. Indeed, the first approach multiplicatively increases all coefficients of the covariance matrices, whereas the second approach only increases their diagonal terms. Hence, the first approach increases the size of target regions while preserving their orientation, whereas the second approach leads to targets that become progressively aligned with the axes, hence losing their orientations (off-diagonal terms becoming negligible compared to increased diagonal terms). It is these changes in orientation that leads to the differences in the "drifting trajectories" of the results obtained by the two approaches.

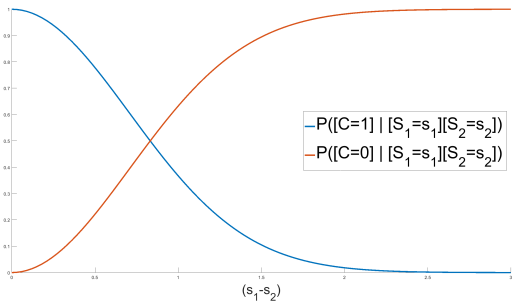

Figure 8.9: Illustration of coherence and anti-coherence constraints. *Eqs (8.11) and (8.12) as functions of* $(s_2 - s_1)$ *in the case of the quadratic measure* $d = (s_2 - s_1)^2$ *(i.e.,* $\eta = \frac{1}{2}$*).*

4 Generalization: From Boolean to continuous coherence variables

We finish this chapter by presenting a preliminary theoretical result about the generalization of the definition of coherence variables. This generalization was motivated by the aim to explore a further alternative approach for weighting sensory pathways in the model.

We have presented a first generalization of the coherence constraint that enables to relax the strict equality requirement and modulate the strength of the corresponding sensory pathway. Yet, the involvement of these constraints are controlled by Boolean coherence variables, acting as switches, either fully activating or fully deactivating the corresponding sensory-matching constraints. Intuitively, one could expect that an alternative approach for the weighting of sensory pathways would be to further generalize coherence variables, so that they would be continuous instead of Boolean, in order to implement progressive activation cursors rather than all-or-nothing switches. In this section we present such generalized continuous coherence variables and explore their effect on the resulting planning process.

4.1 Coherence and anti-coherence

We have defined coherence variables as Boolean variables, with two possible states being either 0 or 1, and we have mainly focused on the $C = 1$ case corresponding to the activation of the sensory-matching constraint, which we write in terms of a general measure d as:

$$P([C = 1] \mid [S_1 = s_1]\ [S_2 = s_2]) \quad = \quad e^{-d(s_1,s_2)}. \tag{8.11}$$

The Boolean switch analogy can be misleading, because it erroneously suggests that value 0 of the coherence variable corresponds to the deactivation of the constraint. This interpretation is incorrect because C can take only two values, 0 or 1, and none of these values can actually deactivate the constraint. Indeed, the sum of $P(C \mid [S_1 = s_1]\ [S_2 = s_2])$ over the two possible values of C must be 1 (due to the normalization rule, Eq (5.3)), and therefore:

$$\begin{aligned} P([C = 0] \mid [S_1 = s_1]\ [S_2 = s_2]) \quad &= \quad 1 - P([C = 1] \mid [S_1 = s_1]\ [S_2 = s_2]) \\ &= \quad 1 - e^{-d(s_1,s_2)}. \end{aligned} \tag{8.12}$$

As illustrated in Fig 8.9, the functions defined by Eqs (8.11) and (8.12) implement two complementary constraints: while Eq (8.11) gives high probability to matching sensory values (small

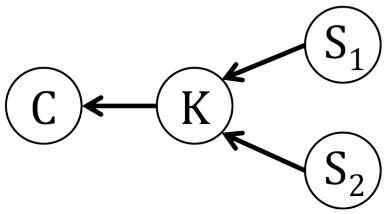

Figure 8.10: Graphical representation of the model defining continuous coherence variables. *The diagram represents the decomposition given by Eq (8.15).*

values of $d(s_1,s_2)$), Eq (8.12) gives high probability to non-matching sensory values (large values of $d(s_1,s_2)$). Hence, while $[C = 1]$ activates a "coherence constraint" with Eq (8.11), $[C = 0]$ activates an "anti-coherence" constraint with Eq (8.12).

Therefore, instead of describing coherence variables as Boolean switches activating or not a coherence constraint, it is more accurate to describe them as Boolean switches activating either a coherence or an anti-coherence constraint. The deactivation of the constraint is implemented by ignoring the value taken by the coherence variable, which leads to a constant unit factor in all computations due to the normalization rule.

4.2 Generalization

To motivate the generalization from Boolean to continuous coherence variables, note that Eqs (8.11) and (8.12) can be written in a compact way as:

$$P([C = c] \mid [S_1 = s_1]\ [S_2 = s_2]) \quad \propto \quad c\ e^{-d(s_1,s_2)} + (1 - c)(1 - e^{-d(s_1,s_2)}). \tag{8.13}$$

It is easy to verify that Eq (8.13) indeed leads to Eqs (8.12) and (8.11) for $c = 0$ and $c = 1$, respectively.

An interesting feature of Eq (8.13) is that it directly provides a continuous family of expressions that generalize coherence variables from Boolean to continuous values. In other words, in addition to compactly formulating constraints for values $[C = 0]$ and $[C = 1]$, Eq (8.13) also defines a constraint for every value between them. In particular, note that for $c = \frac{1}{2}$, Eq (8.13) becomes:

$$\begin{aligned} P([C = c] \mid [S_1 = s_1]\ [S_2 = s_2]) \quad &\propto \quad \frac{1}{2}\ e^{-d(s_1,s_2)} + \frac{1}{2}(1 - e^{-d(s_1,s_2)}) \\ &\propto \quad \frac{1}{2}, \end{aligned} \tag{8.14}$$

which is independent of s_1 and s_2, and hence corresponds to an absence of constraint.

Eq (8.13) can be interpreted as a gradual transformation from the coherence constraint to the anti-coherence constraint, following the increase of a linear parameter. Of course, other alternative non-linear or non-symmetric progressions can also be defined. In order to further generalize and formalize this idea, note that Eq (8.13) can also be interpreted as the result of an inference question involving an additional binary latent variable. Indeed, let us denote this latent variable by K and define a Bayesian model based on all four variables (C, S_1, S_2 and K) with a decomposition of their joint probability distribution represented in Fig 8.10 and given by:

$$P(C\ S_1\ S_2\ K) \quad = \quad P(S_1)P(S_2)P(K \mid S_2\ S_1)P(C \mid K). \tag{8.15}$$

Assuming that $P(S_1)$ and $P(S_2)$ are uniform probability distributions, we can compute $P([C = c] \mid [S_1 = s_1]\ [S_2 = s_2])$:

$$\begin{aligned} P([C = c] \mid [S_1 = s_1]\ [S_2 = s_2]) &\propto \sum_K P([C = c]\ [S_1 = s_1]\ [S_2 = s_2]\ K) \\ &\propto \sum_K P(K \mid [S_1 = s_1]\ [S_2 = s_2])\ P([C = c] \mid K) \\ &\propto f(s_1,s_2)\ g(c) + (1 - f(s_1,s_2))\overline{g}(c) \end{aligned} \tag{8.16}$$

where, for simplicity of notations, we have defined:

$$f(s_1,s_2) = P([K = \text{“Coherence”}] \mid [S_1 = s_1]\ [S_2 = s_2]) \tag{8.17}$$
$$g(c) = P([C = c] \mid [K = \text{“Coherence”}]) \tag{8.18}$$
$$\overline{g}(c) = P([C = c] \mid [K = \text{“Anti-Coherence”}]), \tag{8.19}$$

and we have labeled by {“Coherence”;“AntiCoherence”} the two states of the binary latent variable K.

If we identify $P([K = \text{“Coherence”}] \mid [S_1 = s_1]\ [S_2 = s_2])$ to the definition of the coherence constraint:

$$P([K = \text{“Coherence”}] \mid [S_2 = s_1]\ [S_1 = s_1]) = e^{-d(s_2,s_2)}, \tag{8.20}$$

and

$$P([C = c] \mid [K = \text{“Coherence”}]) = c \tag{8.21}$$
$$P([C = c] \mid [K = \text{“AntiCoherence”}]) = 1 - c, \tag{8.22}$$

we obtain the same expression as in Eq (8.13). However, this time each term in Eq (8.16) can be interpreted in terms of particular probability distributions which can be further generalized. For instance, $P([C = c] \mid K)$ corresponds to the probability of having the cursor at a value c given the state of latent variable K. Eqs (8.21) and (8.22) define these probabilities as linearly decreasing or increasing functions, however they could also be written as sigmoids functions in order to modulate the sensitivity on the displacement of cursor C.

An additional advantage of this formulation is that is suggests a more intuitive definition of the range of values for variable C. Insetad of ranging in $[0, 1]$, we could redefine their range to $[-1, 1]$ such that anti-coherence would correspond to $[C = -1]$, coherence to $[C = 1]$, and deactivation to $[C = 0]$. Eqs (8.21) and (8.22) can be redefined for this new range of values as:

$$P([C = c] \mid [K = \text{“Coherence”}]) = \frac{1 + c}{4} \tag{8.23}$$
$$P([C = c] \mid [K = \text{“AntiCoherence”}]) = \frac{1 - c}{4}, \tag{8.24}$$

where the factor $\frac{1}{4}$ is required for normalization in $[-1, 1]$. Using Eqs (8.21) and (8.22) in Eq (8.16) gives:

$$\begin{aligned} P([C = c] \mid [S_1 = s_1]\ [S_2 = s_2]) &= \frac{1 + c}{4}\ e^{-d(s_1,s_2)} + (\frac{1 - c}{4})(1 - e^{-d(s_1,s_2)}) \\ &= \frac{1}{4}\left(1 - c + 2c\ e^{-\frac{(s_1-s_2)^2}{2\eta^2}}\right), \end{aligned} \tag{8.25}$$

where, on the last line, we have chosen $d(s_1,s_2)$ as being the quadratic measure defined in Section 3.2. We will proceed, in the remaining of this section, with this last expression.

4.3 Effect on the resulting planning process

We have generalized the definition of coherence variables in order to take continuous rather than Boolean values. With this generalization, coherence variables are seen as cursors that can continuously slide from the activation of an anti-coherence constraint at one extreme ($[C = -1]$) to the activation of a coherence constraint at the other extreme ($[C = 1]$), while passing through a deactivating state ($[C = 0]$).

We now evaluate the effect of this generalization on the model, and more specifically on the outcome of motor planning. Details of the computations are provided in Annex B and lead to:

$$\begin{aligned} &P([M = m] \mid \Phi\ [C_A = c_A]\ [C_S = c_S]) \\ &\quad \propto \ (1 - c_A)(1 - c_S)(2\pi)^{-\frac{n_A+n_S}{2}} \\ &\qquad +2\left((1 - c_A)c_S\eta_S(2\pi)^{\frac{-n_A}{2}}\mathcal{N}_S(\rho_s(m) \mid \Phi,\, \eta_S) + (1 - c_S)c_A\eta_A(2\pi)^{\frac{-n_S}{2}}\mathcal{N}_A(\rho_a(m) \mid \Phi,\, \eta_A)\right) \\ &\qquad +4\ c_A\ c_S\ \eta_S\ \eta_A\ \mathcal{N}_S(\rho_s(m) \mid \Phi,\, \eta_S)\ \mathcal{N}_A(\rho_a(m) \mid \Phi,\, \eta_A), \end{aligned} \tag{8.26}$$

where $\mathcal{N}_S(\rho_s(m) \mid \Phi,\, \eta_S)$ and $\mathcal{N}_A(\rho_a(m) \mid \Phi,\, \eta_A)$ are the same normal distributions defined in Eqs (8.9) and (8.10), and n_A and n_S are the number of dimensions of auditory and somatosensory spaces.

Eq (8.26) can be divided into three parts. The term on the first line is independent of m and depends only on values taken by C_A and C_S. On the second line, we find a linear mixture of the two normal distributions, $\mathcal{N}_S(\rho_s(m) \mid \Phi, \eta_S)$ and $\mathcal{N}_A(\rho_a(m) \mid \Phi, \eta_A)$, weighted by coefficients that depend on values taken by C_A and C_S as well as parameters η_S and η_A. Finally, in the last line, we find the same product of terms obtained in Eq (8.8).

Result for extreme values of C_A and C_S The influence of each of the terms in Eq (8.26) is governed by values taken by C_A and C_S. For instance, when both variables take value 0, the second and third lines vanish and the outcome of the planning question is constant (i.e., both constraints are deactivated). When both variables take value 1, the first and second lines vanish and the outcome of the planning question is the multiplicative fusion as obtained in Eq (8.8) (i.e., both constraints are fully activated). Finally, when one of the variables take value 1 and the other 0 (for instance $[C_A = 1]$ and $[C_S = 0]$), all terms vanish, except the normal distribution in the second line corresponding to the coherence variable taking value 1.

In summary, the extremes values of the continuous coherence variables lead to the same planning results as in the previous Boolean switch approach. We now turn to the case of intermediate values of the continuous coherence variables.

Result for intermediate values of C_A and C_S For the sake of illustration, we consider a simplified example where both sensory spaces are one-dimensional[3]. We also consider an instance of the model where somatosensory and auditory planning processes do not agree (as in the case of adaptation in Section 2). Fig 8.11 illustrates the case where the somatosensory

[3]In this case $n_A = n_S = 1$ and Eq (8.26) becomes:

$$\begin{aligned} &P([M = m] \mid \Phi\ [C_A = c_A]\ [C_S = c_S]) \\ &\quad \propto \ (1 - c_A)(1 - c_S) \\ &\qquad +2\sqrt{2\pi}\,((1 - c_A)c_S\ \eta_S\ \mathcal{N}_S(\rho_s(m) \mid \Phi, \eta_S) + (1 - c_S)c_A\ \eta_A\ \mathcal{N}_A(\rho_a(m) \mid \Phi, \eta_A)) \\ &\qquad +8\pi\ c_A\ c_S\ \eta_S\ \eta_A\ \mathcal{N}_S(\rho_s(m) \mid \Phi, \eta_S)\ \mathcal{N}_A(\rho_a(m)|\Phi, \eta_A). \end{aligned} \tag{8.27}$$

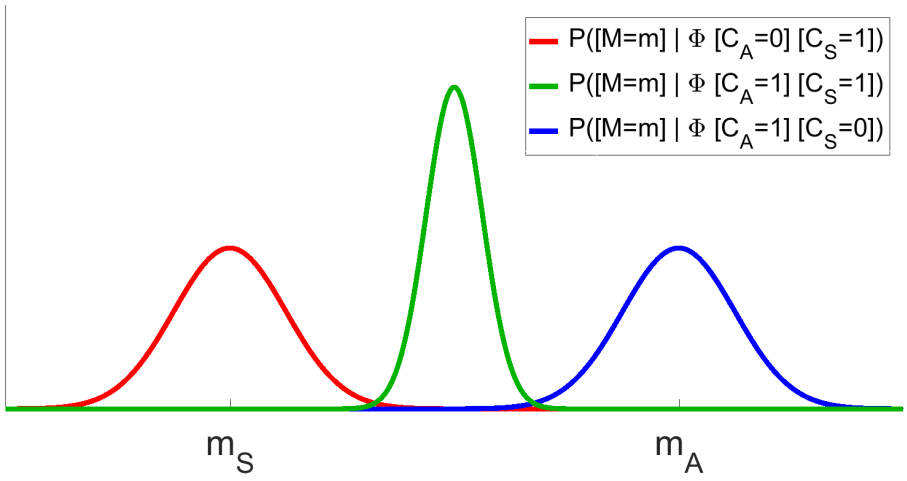

Figure 8.11: One dimensional implementation of the model illustrating the outcome of the three planning processes in the case of adaptation leading to different auditory an somatosensory plannings. *The red, blue and green curves correspond respectively to the outcome of the somatosensory, auditory and fusion plannings.*

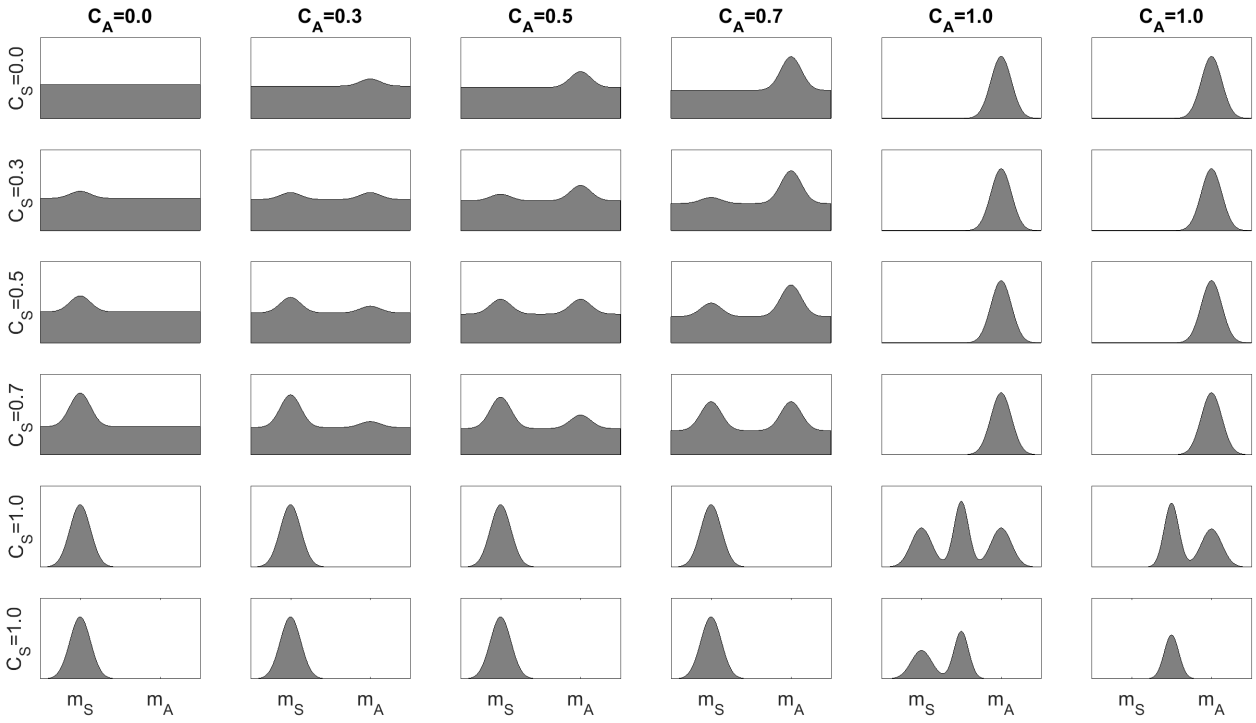

Figure 8.12: Results of the planning process with continuous coherence variables in a simplified one dimensional implementation. *Panels represents the distribution obtained from Eq (8.26) for different values of C_A (columns) and C_S (lines) and fixed value of η_A and η_S ($\eta_A = \eta_S = \frac{1}{15}$).*

planning process (as obtained by $P([M = m] \mid \Phi\ [C_A = 0]\ [C_S = 1])$, red curve in Fig 8.11) results in planned motor commands that are centered around value m_S; the auditory planning process (as obtained by $P([M = m] \mid \Phi\ [C_A = 1]\ [C_S = 0])$, blue curve in Fig 8.11) results in planned motor commands that are centered around value m_A; and finally the fusion planning process (as obtained by $P([M = m] \mid \Phi\ [C_A = 1]\ [C_S = 1])$, green curve in Fig 8.11) results in planned motor commands that are centered halfway between the two previous ones (value $\frac{m_A+m_S}{2}$).

Fig 8.12 illustrates the outcome of motor planning given by Eq (8.26), for different values of C_A and C_S, and fixed values of parameters η_A and η_S. The extreme values of C_A and C_S discussed previously correspond to the four panels on the corners. For small values of C_A or C_S (the first four panels from left to right and from top to bottom), the fusion term of third line in Eq (8.26) is negligible and the planning process corresponds to a mixture of a constant and a bi-modal distributions. However, for values of C_A or C_S closer to 1, the constant term becomes negligible, and the planning results in a mixture of the multiplicative fusion with the two linear terms. The closer C_A and C_S get to 1, the smaller the contribution of the linear terms, and thus the planning process converges to the single multiplicative fusion when both C_A and C_S take value 1.

4.4 Discussion

In summary, contrary to what we intuitively expected, the generalization to continuous coherence variables does not result in the modulation of the contribution of each sensory pathway directly, as in the previous two approaches. Indeed, intermediate values of coherence variables do not shift the location of the resulting fusion process from one sensory planning mode to the other. Instead, continuous coherence variables enable to shift between two main forms of planning: 1) one in which both sensory pathways are combined by fusion (when $C_A = C_S = 1$); 2) one in which both sensory pathways are considered alternatively, resulting in a bi-modal distribution. In this second configuration, the probability of drawing from one or the other mode depends on the relative values of C_A and C_S: when C_A is greater that C_S, the auditory mode becomes predominant, and *vice versa*.

Parameters η_A and η_S control how sensitive the outcome of the planning process is with respect to changes in C_A and C_S. Fig 8.13 illustrate the result of motor planning for the same values of C_A and C_S as in Fig 8.12, but for different values of parameters η_A and η_S.

Finally, notice that the intermediate bi-modal planning could only be obtained by considering the soft constraint implemented with parameters η_A and η_S. Indeed, when η_A and η_S tend towards 0 (leading to the hard constraint) the results of the planning process jumps discontinuously from the multiplicative fusion when $C_A = C_S = 1$, to the constant distribution (i.e., no constraint) when C_A and C_S differ from 1, as illustrated in Fig 8.13 for parameters $\eta_A = \eta_S = 10^{-9}$.

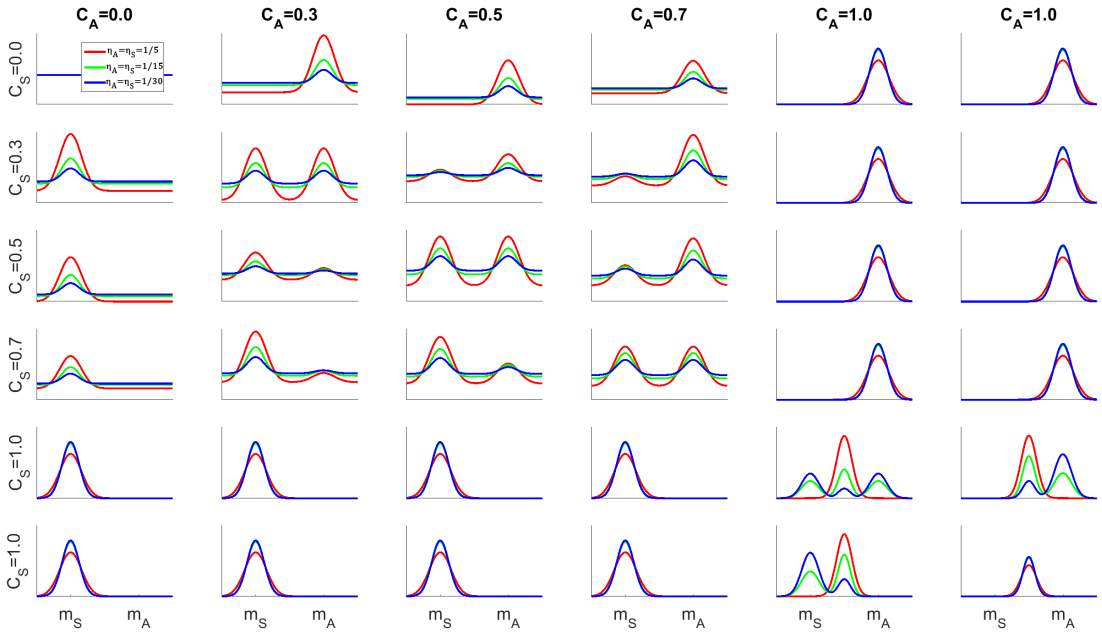

Figure 8.13: Illustration of the sensitivity of the planning process with continuous coherence variables with respect to three values of parameters η_A and η_S. Curves represent the distributions obtained from Eq (8.26) for the same values of C_A and C_S as in Fig 8.12, with three different values of parameters $\eta_A = \eta_S = \{\frac{1}{5}, \frac{1}{15}, \frac{1}{30}\}$, corresponding respectively to red, green and blue curves.

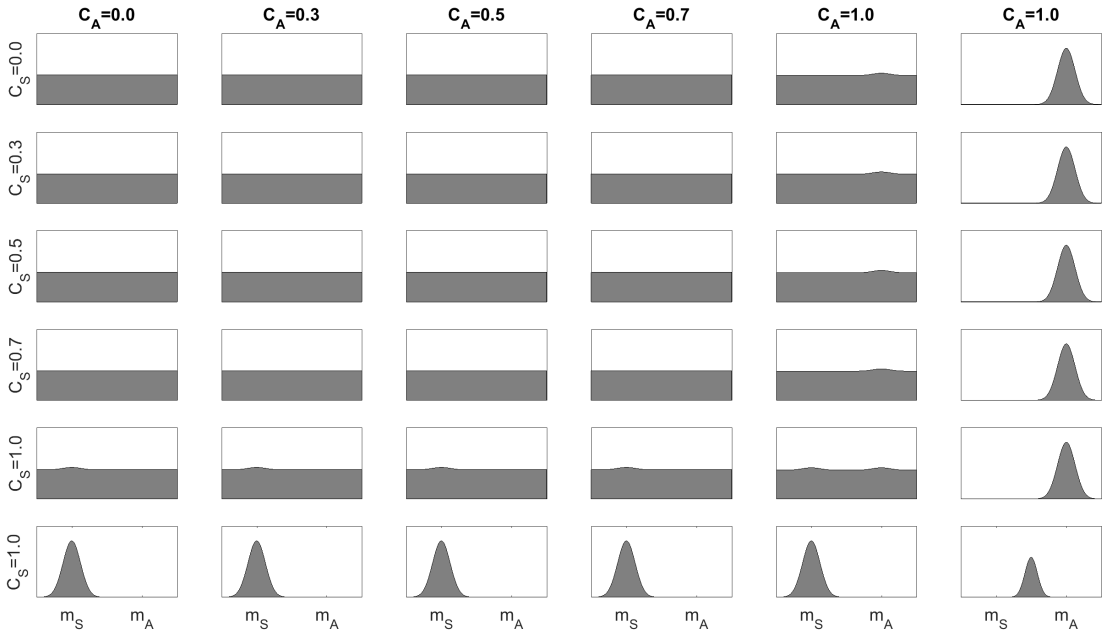

Figure 8.14: Absence of intermediate bi-modal planning when values of parameters η_A and η_S tend toward 0. *Curves represent the distributions obtained from Eq (8.26) for the same values of C_A and C_S as in Fig 8.12 but for values of parameters $\eta_A = \eta_S = 10^{-9}$.*

Chapter 9

Experimental studies

1 Introduction

One of the aims of our modeling work is to explore the involvement of somatosensory information in speech motor planning. In previous chapters, we have developed a model for speech motor planning based on the central hypothesis that phonemic goals are characterized both in auditory and somatosensory terms.

We have conducted two experimental studies in order to further examine the characteristics of somatosensory goals [4]. These studies were designed to evaluate two assumptions of our modeling work. Firstly, as compared to the somatosensory correlates of auditory targets, somatosensory targets correspond to a more local (prototypical) description of articulatory configurations; the aim of the first experiment, presented in Section 2, is to evaluate this hypothesis based on a measure of articulatory variability in presence versus absence of auditory feedback. Secondly, somatosensory targets should enable to perform phoneme identification based on somatosensory information alone; the aim of the second experiment, presented in Section 3, is to explore if this is indeed the case and evaluate the efficiency of this process.

2 Auditory versus somatosensory control of speech: is token-to-token variability influenced by the presence or absence of auditory feedback?

2.1 Goals and motivations

This first study aims at evaluating the hypothesis that somatosensory targets would correspond to a more prototypical description of articulatory configurations, as compared to the full set of articulatory configurations enabled by auditory targets. In particular, we expect that when speech relies only on somatosensory inputs (absence of auditory feedback) the resulting articulatory variability would reveal the domain of variation allowed by somatosensory targets. Hence, if somatosensory targets are indeed more prototypical, we expect that, in the absence of auditory feedback, articulatory variability would be reduced. In order to test this prediction, we compared token-to-token variability in the articulatory domain associated with speech production in the presence versus absence of auditory feedback. In order to specify the design of our experimental setup, we need to address two particular issues. Firstly, how to remove auditory feedback? This will inform our choice of experimental conditions. Secondly, what speech items to use? This will inform our choice of experimental corpus.

[4]We thank Pamela Trudeau-Fisette and Christophe Savariaux for their crucial help in designing and conducting the experiment, David Ostry for his help in designing the experiment and analyzing the results, and Silvain Gerber for performing the statistical analyses.

2.1.1 Choice of conditions: Removing auditory feedback

We aim at comparing speech production in the presence and absence of auditory feedback. However, removing auditory feedback is not that straightforward. A first possibility would be to compare normal speech with unvoiced speech. However, unvoiced speech is not just speech with an absence of sound. It is likely that unvoiced speech relies on a different control strategy, since it requires activating different muscles; this may impact the pattern of articulated gestures. Indeed, in agreement with results of other studies (Crevier-Buchman et al., 2011; Dromey & Black, 2017; Hueber, Badin, Savariaux, Vilain, & Bailly, 2010; Janke, Wand, & Schultz, 2010), our pilot measurements suggested important articulatory differences between voiced and unvoiced speech. We interpreted these differences as resulting from contracting muscles in the pharyngeal region during voicing, which may mechanically impact tongue shapes. A second possibility would be to compare normal speech with normal speech with auditory feedback masked with noise. However, while masking noise can certainly mask air conducted sounds, bone conduction is difficult to mask with comfortable levels of noise.

The solution that we finally adopted was to compare whispering speech (W condition) with whispering speech with auditory feedback masked by an 80 dB(A) noise (WN condition). The level of masking noise was chosen as sufficiently high in order to mask whispered speech, while being comfortable for the subject. All subjects reported not hearing themselves while whispering. In order to reduce possible hyperarticulation or intensity variation induced by the masking noise (the so-called "Lombard effect" (Castellanos, Benedí, & Casacuberta, 1996; Egan, 1972; Junqua, 1993; Lane & Tranel, 1971; Lombard, 1911)), visual feedback was provided (through a level meter displayed on the screen). In addition, we further expected that using a bite block would reduce the trend for hyperarticulation usually associated with wide jaw opening. In order to assess how the whispering condition (W) can be compared with normal speech, we further included a third condition consisting of normally voiced speech (N).

Each subject performed the three conditions (N, W and WN) during the experimental session. This choice raises an additional problem however, since the order in which subjects perform conditions may influence their production precision. Indeed, the variability of productions on the condition performed first may differ from the condition performed last, because of learning, habituation or boredom. In order to counterbalance the influence of this potential order effect, two orderings were selected by permuting the first and last conditions (N-W-WN and WN-W-N). We did not include the additional 4 possible permutations of conditions due to limitations on the number of subjects. Our main focus is on the presence or absence of auditory feedback, hence we just consider the two ordering groups where only the whispering condition with masking noise was presented at the beginning or at the end, and the whispering condition without masking-noise was unchanged.

2.1.2 Choice of corpus: CV-syllables

Our study focuses on the production of isolated vowels. This choice aims at keeping the task simple, while controlling and minimizing at best other possible sources of variability. In order to control and minimize variability arising from coarticulation, we focus on the production of CV syllables, with a unique initial /l/ consonant. In addition, subjects were provided with a visual feedback of their tongue posture in order to begin utterances with the same static /l/ posture. The production of consonant /l/ requires a high jaw position and a high tongue tip position, with some freedom in the anterio-posterior position of the tongue and in the elevation of the back part of the tongue. Thus, starting with this consonant before producing the vowel was considered as a good strategy to incite speakers to start movement from the same region of

the vocal tract, without strongly influencing the articulation of the subsequent vowel, as would have done alveolar fricatives /s, z/ or stops /t, d/.

Furthermore, we focus only on tongue control and therefore consider a set of five vowels (/e/, /ɛ/, /a/, /ɔ/, /œ/) that do not involve lip movements. In order to maintain a certain level of naturalness, the CV syllables were presented as lexical French words or parts of lexical French words: "les" (plural "the") for /le/, "lait" ("milk") for /lɛ/, "la" (feminine "the") for /la/, "lors" ("during") for /lɔ/ and "leur" ("their") for /lœ/.

Finally, the influence of jaw control on tongue control was removed by means of a bite block specifically molded for the teeth of each subject. The bite block (Coltene PRESIDENT putty) was molded by the subject while holding a reference of 15 mm between their teeth (see right panel of Fig 9.5 for illustration). This size, quite large as compared to usual bite block experiments, was motivated by the wish to keep the same bite block for experiments 1 and 2, and by the necessity to give enough freedom to tongue movements in experiment 2.

2.2 Materials and Methods

The experimental protocol was performed in accordance with the ethical standards specified by the 1964 declaration of Helsinki and approved by the institutional ethics committee of the University Grenoble-Alpes.

2.2.1 Participants

8 subjects (4 males, ages ranging between 20 and 36 years, average 24 years) took part in this experiment. Subjects gave their informed consent before participating and were compensated with 15€ gift cards. Subjects were French native speakers and reported having no cognitive impairment. Subjects showed normal hearing (pure-tone detection thresholds below 25 dB HL at every frequency tested: 250 Hz, 500 Hz, 750 Hz, 1000 Hz, 2000 Hz, 4000 Hz, and 8000 Hz). Subjects were randomly assigned to one of two condition ordering groups (N-W-WN and WN-W-N).

2.2.2 Experimental setup

Subjects were seated in a sound proof room in front of a computer screen placed at one meter in front of them, as illustrated in Fig 9.1. We measured tongue postures by means of an electromagnetic articulography system (EMA). This high temporal resolution three-dimensional motion tracking system is a reliable device for the observation of speech articulators (Savariaux, Badin, Samson, & Gerber, 2017). However, one drawback of this instrument is that it is invasive. It involves gluing sensors connected with wires on the tongue of the subject, which may certainly perturb their production. If the EMA device increases variability, it may hide differences between conditions. Hence, in order to reduce possible influence of the EMA sensors on speech production, we further included a habituation task consisting of the production of 10 phonetically balanced sentences. In addition, in order to assess the impact of EMA sensors on articulatory variability, we also included a preliminary task (reference task) during which subjects were instructed to produce the CV-syllable corpus without sensors.

2.2.3 Experimental design

The experimental session lasted for around 1.5 hours and was composed of two main parts, a preliminary phase and a production phase, as illustrated by the block diagram of Fig 9.3. The

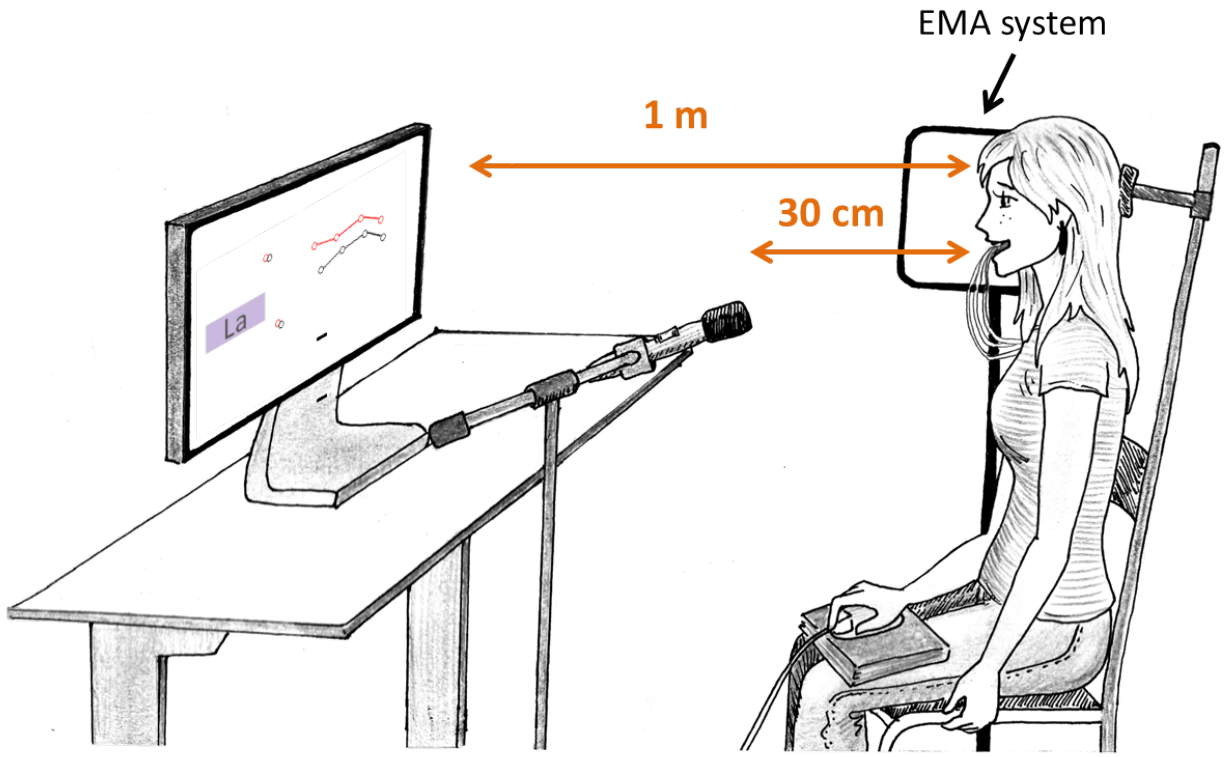

Figure 9.1: Experimental setup. *The computer screen displays the first part of the production task (see Fig 9.4). The computer mouse was used only for categorization answers during the second experiment (see Section 3).*

preliminary phase included three tasks intended for preliminary measures and subject familiarization. The production phase corresponded to the production tasks in the three experimental conditions.

Preliminary phase

Phonetically balanced sentences

Il se garantira du froid avec ce bon capuchon.
Annie s'ennuie loin de mes parents.
Les deux camions se sont heurtés de face.
Un loup s'est jeté immédiatement sur la petite chèvre.
Dès que le tambour bat les gens accourent.
Mon père m'a donné l'autorisation.
Vous poussez des cris de colère?
Ce petit canard apprend à nager.
La voiture s'est arrêtée au feu rouge.
La vaisselle propre est mise sur l'évier.

Figure 9.2: List of phonetically balanced sentences (Combescure, 1981).

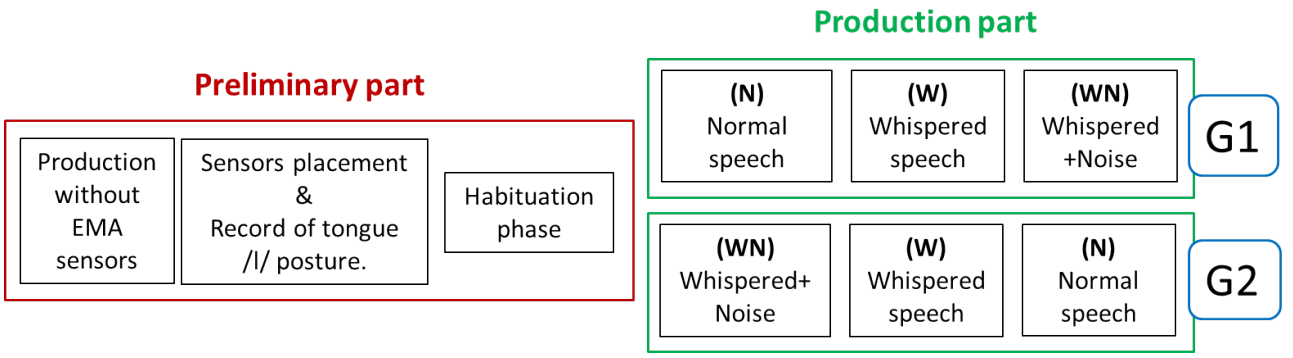

Figure 9.3: Block diagram of the experimental design.

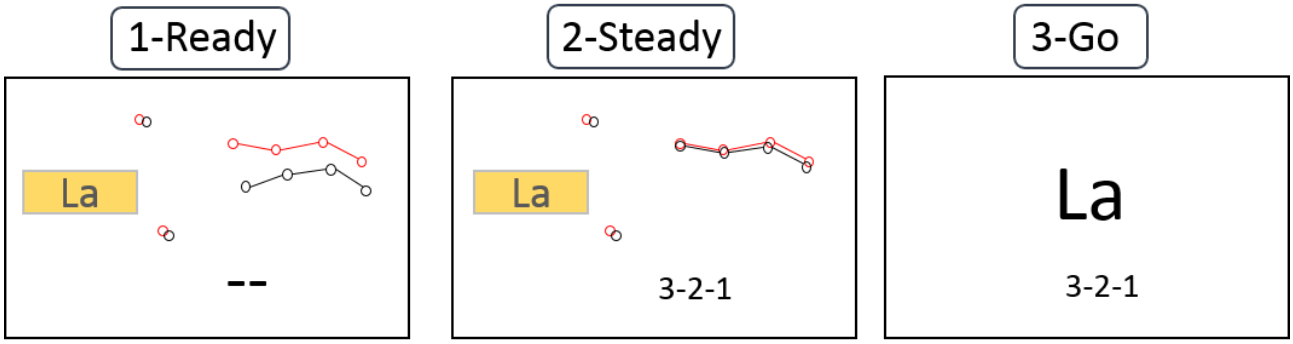

Figure 9.4: Task Design. *The task was composed of three parts. (1-Ready) Subjects began by indicating that they were ready by positioning their tongue (black line) in the recorded /l/ configuration (red line). (2-Steady) A 3-2-1 countdown was displayed in order to prepare subjects. (3-Go) Immediately after, subjects produced the required syllable while holding the vowel for the duration specified by a second 3-2-1 countdown. The syllable to produce was always presented as a word ("Le", "Lait", "La", "Lors", "Leur") but subjects were instructed to produce only the onset and nucleus.*

Reference task – Production without EMA sensors: The aim of this first task was to familiarize subjects with the design of the task, and to record their productions in normal conditions, without EMA sensors and without the bite block. The words containing CV syllables were displayed on the computer screen along with a 3-2-1 countdown intended to prepare subjects. Subjects were instructed to vocalize the CV syllables with the vowel hold for around 1 second (duration displayed with a second countdown on the screen). Each of the 5 CV items was presented 5 times in a pseudorandom sequence composed of 5 permutations cycles of the 5 items. Subjects performed this reference task in normal and whispered conditions.

Articulatory recording of reference /l/ tongue posture: After placement of EMA sensors and of the bite block, the subject tongue posture for consonant /l/ was recorded. Recording was performed by instructing subjects to position their tongue in a /l/ like configuration that felt as natural as possible for them.

Habituation task – Phonetically balanced sentences: Subjects were instructed to read a set of 10 phonetically balanced sentences (see Fig 9.2) in order to familiarize them with

speaking with the sensors glued to their tongue . Sentences were progressively displayed on the computer screen, and subjects were instructed to read them as naturally as possible.

Production phase The second phase was composed of 3 parts corresponding to the three experimental conditions:

1. **(N)** normal speech with auditory feedback,
2. **(W)** whispered speech with auditory feedback,
3. **(WN)** whispered speech without auditory feedback (masked with noise).

In all speaking conditions, subjects were asked to produce the same 5 CV syllables as during the "production without EMA sensors" task. Each of the 5 CV items was presented 10 times in a pseudorandom sequence composed of 10 permutations cycles of the 5 items. All utterances were produced with the bite block.

Task design:

1. **Ready**: Prior to each utterance, subjects were asked to position their tongue in their recorded /l/ configuration. In order to help them positioning their tongue in the same initial configuration, the recorded /l/ configuration was displayed on the screen along with the real time position of their sensors (see Fig 9.4 left panel).
2. **Steady**: In order to avoid errors and warn the subject of the beginning of the next utterance, a 3-2-1 countdown (lasting around 1 second) was displayed on the screen, along with the identity of the CV syllable to produce (see Fig 9.4 middle panel). Subjects were instructed to hold the /l/ posture (that remained displayed on screen) and begin the utterance at the end of the countdown.
3. **Go**: Immediately after the first countdown, tongue display disappeared and a second 3-2-1 countdown began (lasting around 1.5 seconds) in order to provide a reference for the holding duration of the vowel (see Fig 9.4, right panel). Subjects were instructed to normally produce the initial /l/ but hold the vowel for the duration of the countdown. Only during the (WN) condition, auditory feedback was masked with the 80 dB(A) white noise sent through in-ear headphones (compatible with the EMA system, Etymotic ER1). In-ear headphones were not used during the other two conditions.

2.2.4 Recordings and data processing

Articulatory data It was collected at 200 Hz via an electromagnetic articulography system (EMA Wave-NDI). The recorded articulatory data was filtered at 30 Hz. In order to observe tongue movements, 4 EMA sensors were placed midsagittaly on the tongue: one on the tongue tip, two on the tongue body and one on the tongue dorsum (see Fig 9.5). Two additional EMA sensors were placed on the upper and lower lip in order to track and prevent possible lip movements. Subjects were provided with real time visual feedback of these sensors and were asked to keep them as stationary as possible during the experiment. Finally, the reference EMA sensor, with respect to which the positions of other sensors are defined, was placed on the nasion. Fig 9.5 shows the location of each of these 7 EMA sensors. Sensors were placed on the tongue with PeriAcryl 90HV, a high viscosity Cyanoacrylate Oral Adhesive safe to use on oral tissues.

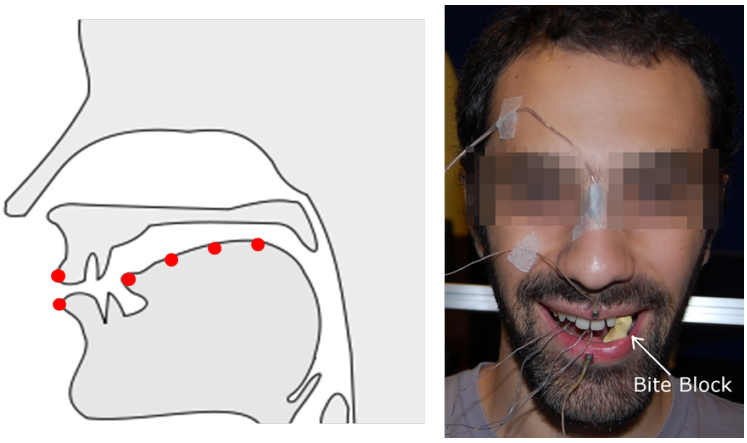

Figure 9.5: Sagittal view of sensors locations (left panel) and frontal view of bite block (right panel).

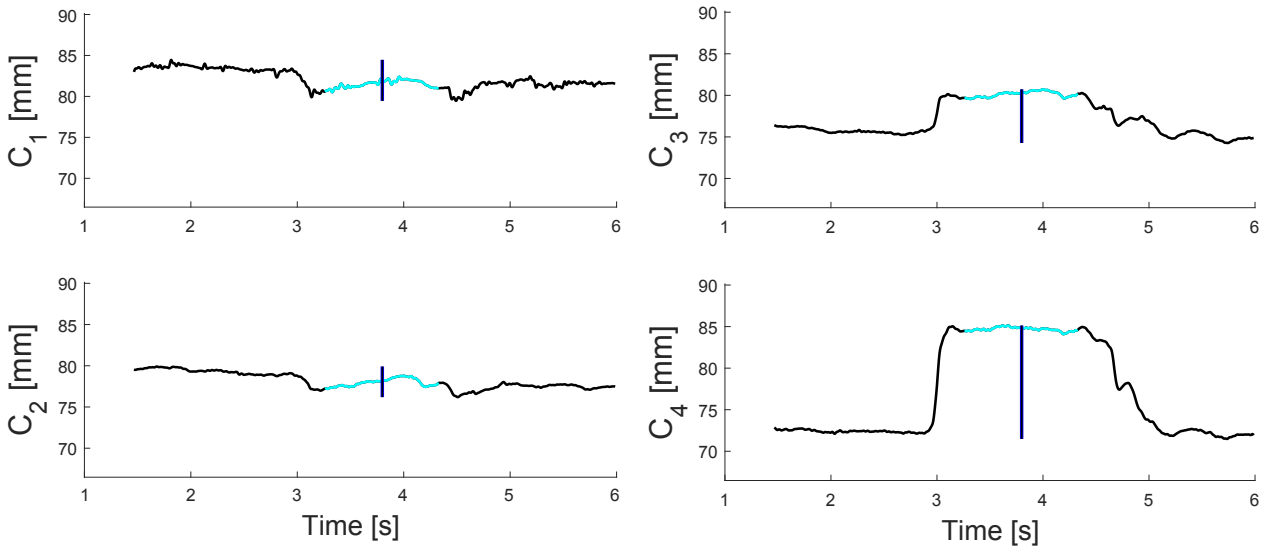

Figure 9.6: Example of articulatory data selection. *Time evolution of each sensor position; the best stability window is indicated by the cyan portion of the curves. The vertical blue line indicates the instant of the retained tongue posture (center of the stability window).*

For each utterance, a representative tongue posture was manually selected by identifying a window of best articulatory stability and retaining the tongue posture at its center. This selection was performed with a custom visualization program (*Matlab*), as illustrated in Fig 9.6. The four panels represent the time evolution of the norm of the 3-dimensional vector position of the four EMA sensors. The window of best stability was selected with respect to the sensor that presented the clearest transition, which generally depended on the utterance. In the example given in Fig 9.6, the best stability window was selected with respect to the fourth sensor (bottom right panel). The duration of the selected window varied from around 1 to 1.5 seconds, depending on the holding time maintained by the subject.

Acoustic data In order to assess the influence of coils on the variability of productions we analyze the token-to-token variability of vocalization produced during the preliminary task

(without coils and without bite bloc) and during the normal condition of the production task. Acoustic signal was captured through a microphone (AKG C1000S) and recorded with Audacity software at 44100 Hz using a Tascam US 600 sound card. The microphone was placed 30 cm in front of the subject (see Fig 9.1). The recorded signals was subsequently down-sampled to 8 kHz and the first two formant frequencies were estimated using the linear predictive coding (LPC) algorithm (*Matlab*).

2.2.5 Statistical analysis

Articulatory data Our aim is to assess whether the speaking condition (Normal (N), Whispering (W) and Whispering with Noise (WN)) influences production accuracy. We quantify production accuracy by a dispersion measure , for each of the 4 sensors, in the context of each of the five vowels. The dispersion measure is computed as the average distance of the sensor to the average position of all repetitions of the considered vowel[5]. In addition to the effect of the condition on production accuracy, we also take into account possible effects of the vowel and of sensor number.

Since collected data correspond to several measurements performed for each subject, the independence of observations is not guaranteed, which is a necessary condition for the application of classical ANOVA. Hence, in order to take into account measurement dependencies, we use a linear mixed model, with subject as random effect, performed with the `lme` function of the `lme` package of the R software. Furthermore, since sensor measurements are taken simultaneously and are constrained by tongue structure, their dispersion measures appear to be correlated, which requires further modeling (through the weight and correlation parameters of the `lme` function) the covariance matrix of the model (Bazzoli, Letué, & Martinez, 2015).

We consider subject identity (8 subjects) as the random effect and four fixed effects, which are speaking condition (3 levels), vowel identity (5 levels), sensor number (4 levels) and group (2 levels, corresponding to the order of speaking conditions). The best fitting model was defined following a backward deletion approach (Mundry & Nunn, 2009) using likelihood-ratio test in order to assess the influence of each factor and each interaction between factors. The applicability of the model assumptions was performed by graphical inspection of the residuals.

Once the model is defined, multiple comparison tests are performed following the method presented by Hothorn, Bretz, and Westfall (2008).

Auditory data We performed two control analysis in order to assess, first, whether noise influenced whispering intensity (in order to exclude involvement of Lombard effect (Egan, 1972; Lane & Tranel, 1971; Lombard, 1911)) and, second, whether the perturbations induced by the experimental setup (tongue sensors and bite block) influenced overall production variability. We addressed the first point by assessing the overlap of confidence intervals of the average intensity of whispers (computed as the root mean square of signal) in conditions with and without noise. Confidence intervals were obtained as bootstrapping estimate of 95% confidence errors. Results are presented in Section 2.3

We addressed the second point by analyzing production dispersion during the preliminary task and during the normal condition of the production task. We assessed production dispersion

[5]The dispersion measure σ_i^ϕ, for sensor i, in the context of phoneme ϕ, is given by: $\sigma_i^\phi = \frac{1}{N}\sum_{k=1}^{N} d_{ik}^\phi$, where N is the number of repetitions of each phoneme (10 in our case), and $d_{ik}^\phi = \sqrt{\left(x_{ik}^\phi - \mu_{xi}^\phi\right)^2 + \left(y_{ik}^\phi - \mu_{yi}^\phi\right)^2 + \left(z_{ik}^\phi - \mu_{zi}^\phi\right)^2}$ is the Euclidean distance (in 3-dimensional space) of sensor i towards its average position $\mu_i^\phi = \left(\mu_{xi}^\phi, \mu_{yi}^\phi, \mu_{zi}^\phi\right)$ over the 10 repetitions of phoneme ϕ.

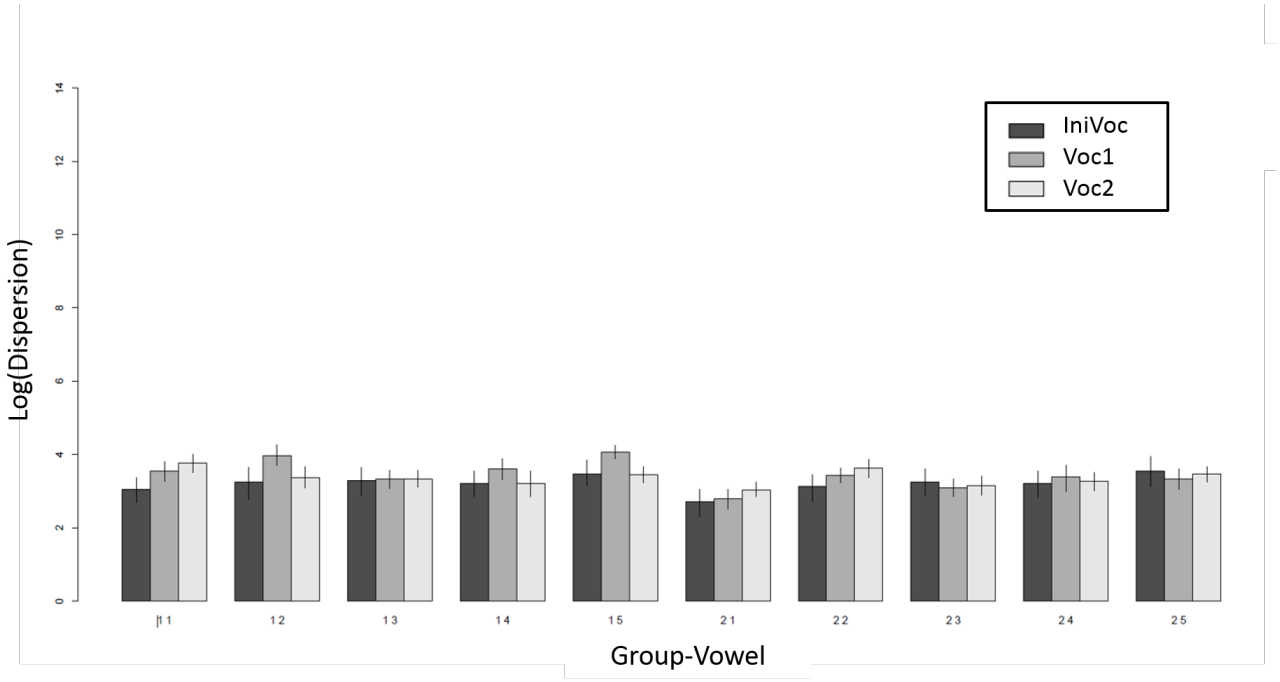

Figure 9.7: Dispersion measures of productions with and without sensors. *Condition "IniVoc" corresponds to productions during the preliminary task (without sensors). Voc1 corresponds to productions during the normal condition of the production task. Voc2 corresponds to the same task as Voc1 but performed on the session of the following day (see Section 3).*

for each vowel and each condition as the Euclidean distance (in two first formant space) of each utterance with respect to the average of utterances of this vowel in this condition. The effect of condition (without sensors *vs* with sensors) was assessed with a linear mixed model with subject as random effect and group, condition, and vowel as fixed effects. Accuracy of the model was achieved by using the logarithm of the dispersion measure as response variable.

2.3 First results

We begin by presenting results of the two control measurements, aiming at confirming that EMA sensors did not induce additional significant variability to production, and at confirming that masking noise did not influence the whispering intensity. Next, we present results corresponding to the main questions addressed by the study: the effect of speaking condition (N, W, WN) on average tongue position and on articulatory variability.

2.3.1 Effect of EMA sensors on productions

In order to confirm that EMA sensors did not induce additional production variability, we analyzed the acoustic dispersion of utterances produced with and without sensors with a linear mixed model, as described in Section 2.2.5. The best fitting model indicated no interaction between group and other factors, the only fixed effects retained were vowel and condition, and their interaction. Comparing the effect of condition on the dispersion measure indicated no significant difference.

Fig 9.7 shows average dispersion measures for each vowel and each group for conditions without sensors (IniVoc) and with sensors (Voc1). Dispersion measures obtained during vocal-

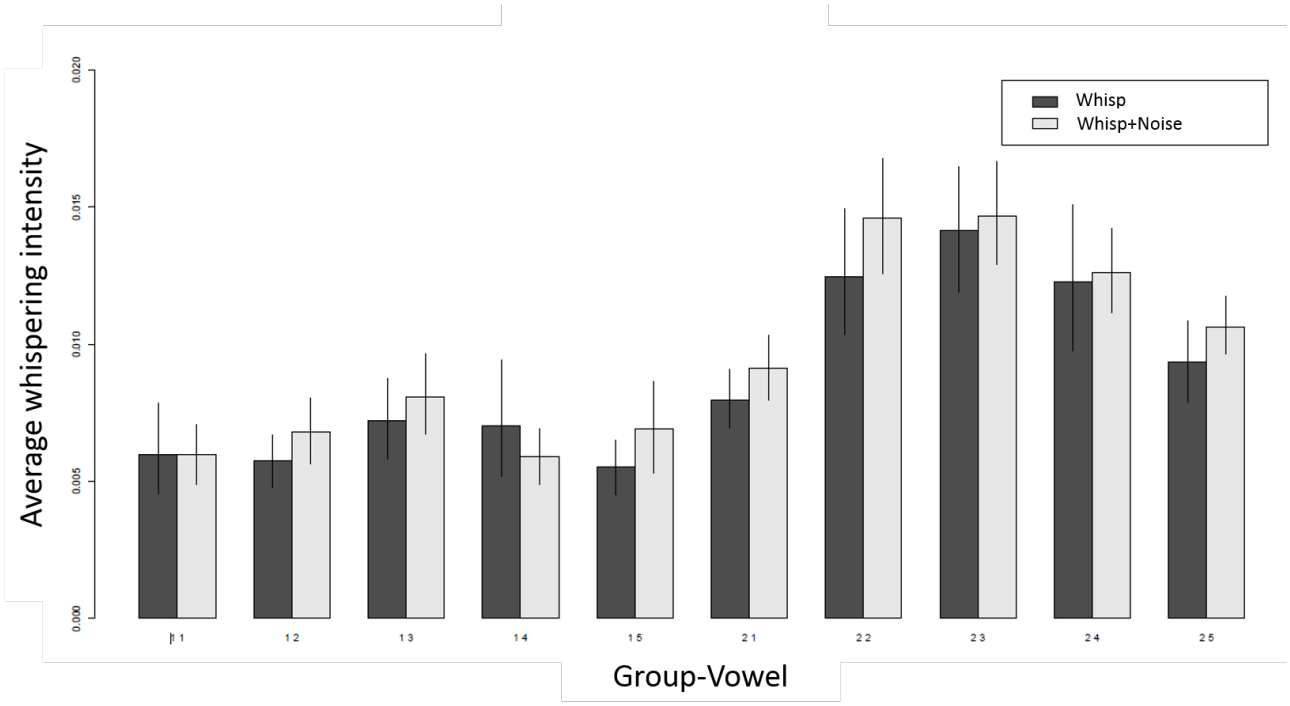

Figure 9.8: Comparison of average whispering intensity. *Bars indicate bootstrapping estimate of* 95% *confidence errors, for each vowel and each group.*

izations with sensors on the following day (next study, see Section 3) are also indicated.

2.3.2 Effect of masking noise on whispering intensity

In order to confirm that masking noise did not induce hyperarticulation effects, we compared the intensity of whispering in conditions with and without noise. Whispering intensity was assessed as the root mean square of the recorded acoustic signal. Fig 9.8 presents the average whispering intensity for each vowel in both whispering and whispering with noise conditions, along with bootstrapping estimate of 95% confidence errors. Confidence intervals of both conditions appear to overlap for all vowels in both groups, indicating that whispering intensity is not influenced by the addition of noise in the auditory feedback. This indicates that productions were not overly influenced by the Lombard effect (Egan, 1972; Lane & Tranel, 1971; Lombard, 1911). In addition, there is a general trend for Group 2 to whisper louder than Group 1.

2.3.3 Effect of speaking condition on articulatory postures

Observing, for each vowel, the average tongue shapes and their variability (Fig 9.9, for illustration) indicates that for all subject except one (subject 3), there does not appear to be any difference in average tongue shapes across condition. Subject 3, on the other hand, has noticeably different average tongue shapes, for vowels /a/, /œ/ and /ɔ/, between the WN condition and the other conditions.

2.3.4 Effect of speaking condition on articulatory variability

The best fitting linear mixed model included significant subject-vowel and subject condition interactions ($p<0.0001$ in both cases). It also included significant interactions between the follow-

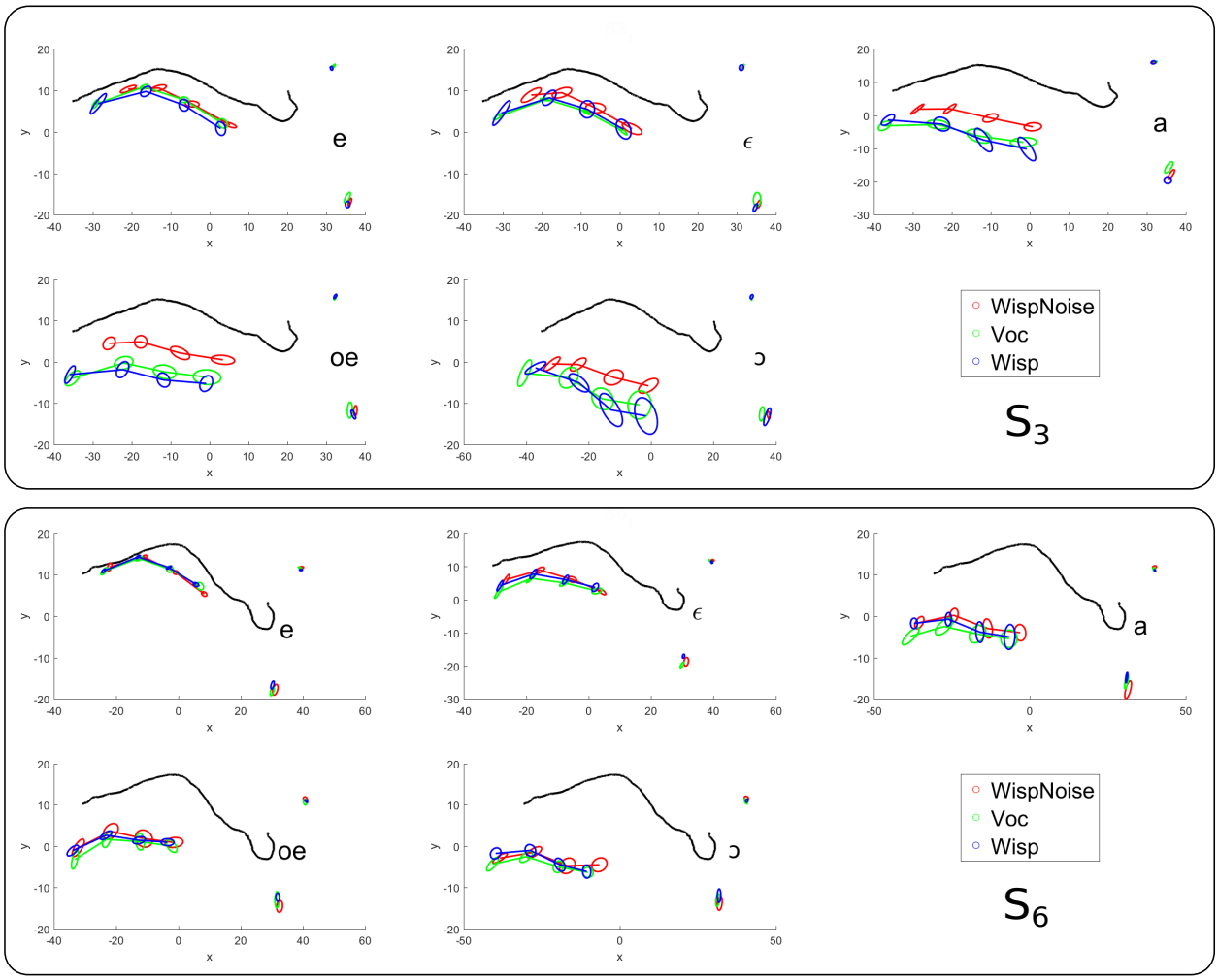

Figure 9.9: Illustration of average productions and dispersion ellipses for each vowel in the three conditions. *The top panel illustrates the general pattern obtained for most subjects (see Annex C for other subjects). The bottom panel shows noticeable differences in mean tongue postures for Subject 3.*

ing fixed effects: condition-sensor ($p=0.05$); sensor-vowel ($p<0.0001$), sensor-group ($p<0.0001$), and condition-group ($p<0.01$).

Closer inspection of the data (see Fig 9.10) reveals that the condition-group interaction corresponds to an absence of significant differences across conditions in group 2, and large significant differences for group 1. We focus now on the results of group 1. For group 1, significant differences between conditions W-WN are systematically found for all sensors and all vowels: the articulatory variability in condition WN is smaller than in condition W. The condition-sensor interaction corresponds to the fact that there is a quasi-systematic significant difference between condition (N) and (W) for the anterior sensors 4 and 3 (larger articulatory variability for condition W) and a quasi-systematic absence of difference for sensors 2 and 1.

2.4 Discussion

We assessed the precision of speaking with or without auditory feedback by evaluating the difference between whispering and whispering with noise conditions. Whispering was chosen

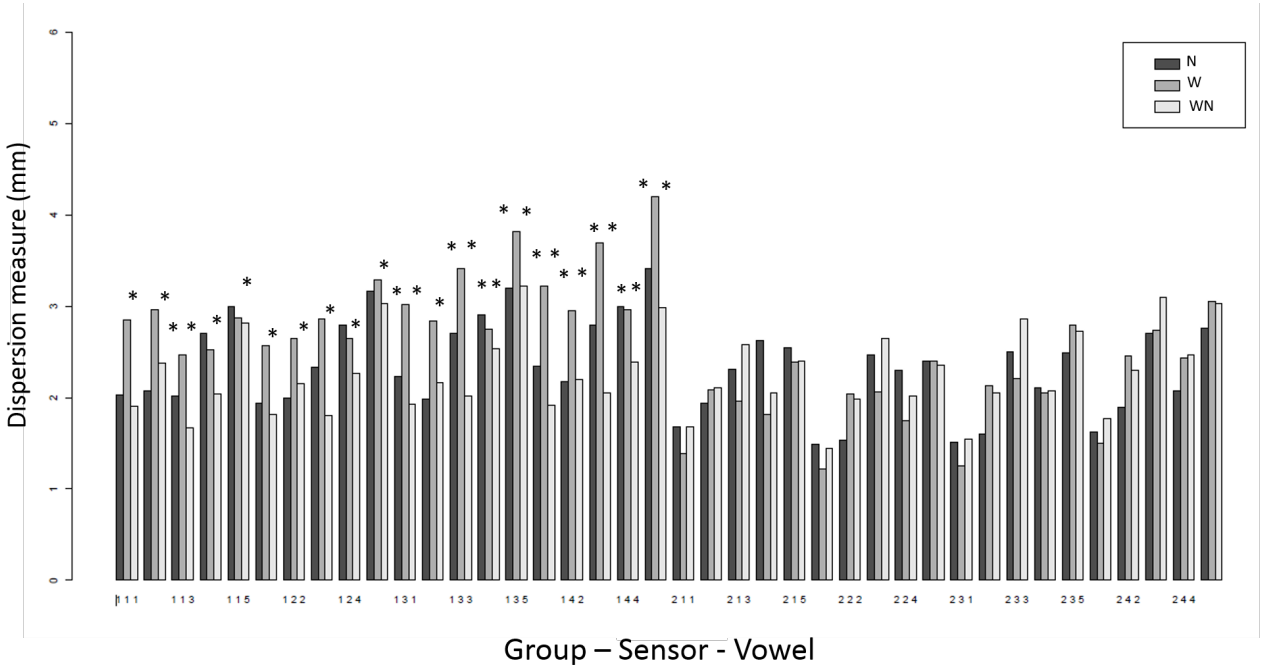

Figure 9.10: Multiple comparison between conditions for each group, sensor and vowel. *Stars indicate significant differences between neighboring bars predicted by the model.*

in order to avoid a possible influence of auditory bone conduction in the noise condition. Furthermore, we performed control measurements in order to confirm two assumptions. Firstly, EMA sensors fixed on subject tongues, while being intrusive, did not significantly affect their production variability. Second, masking noise did not induce hyperarticulation. While more subjects would be needed before drawing definite conclusions, we discuss some interpretations of these preliminary results.

2.4.1 Interpretation of results

Our results indicate a clear difference between conditions WN and condition W in group 1, with condition WN being systematically less variable than condition W. This tends to support our hypothesis of smaller articulatory variability when speech production relies only on somatosensory inputs. However, the fact that this pattern is featured only in the first group, who performed condition WN last (after conditions W and N), could suggest an alternative interpretation, in which the reduction in articulatory variability would result from a learning effect.

However, two observations tends to undermine this alternative explanation. Firstly, in group 2, there is no significant reduction in variability from the first condition (WN) to the last one (W). Secondly, the variability observed in the WN condition in group 2 is similar to the variability observed in the WN condition in group 1 (see Fig 9.10). This suggests that that the small observed variability is an intrinsic property of the WN condition and not the result of a learning effect. In this context, the absence of significant differences across conditions observed for group 2 could be interpreted as a memory effect resulting from condition WN being first, which may expose subjects to prototypical productions, affecting their productions in the following N and W conditions, inducing a limitation of their variability.

We have attributed the reduced variability of condition WN to the absence of auditory

feedback. However, it may also be that it is not the absence of auditory feedback but the presence of noise that influences articulatory variability: in adverse conditions (noisy environment), subjects would be more prototypical in order to improve intelligibility. While this interpretation cannot be fully discarded, it finds little support in our results for two reasons. First, our control measurements indicate that noise did not induce the Lombard effect (Egan, 1972; Lane & Tranel, 1971; Lombard, 1911). Second, since whispers are more difficult to understand, we would also expect whispers to be less variable than in normal condition; this is not supported by our data.

Finally, our interpretation crucially rests on the existence of somatosensory goals in speech. Note, however, that the involvement of somatosensory information in speech production does not necessarily imply the existence of somatosensory goals in speech. Indeed, speech gestures may be essentially planned and controlled with respect to auditory goals, with somatosensory information being used in order to further estimate auditory states via an internal model predicting auditory consequences from somatosensory signals. However, if that were true, we would expect that speaking without auditory feedback would be less accurate than speaking with both auditory and somatosensory feedback, since in the first case the processed auditory states would only be estimations of the real ones.

From all these different points, we conclude that our experimental observations are consistent with the existence of somatosensory goals in speech production and with the hypothesis that these goals would be characterized by more prototypical articulatory patterns as compared with the full set of articulatory configurations enabled by auditory goals.

2.4.2 Caveats and future directions

Since the present study is still preliminary, there are additional questions and analyses that deserve further inspection. First, the origin of the difference between groups certainly needs to be clarified. For this, additional subjects are necessary in order to determine whether the observed difference indeed results from the condition order or just by chance. We could have avoided the problem of ordering condition with two other possible protocols. The first possibility would have been to define three groups assigned to single conditions. We discarded the first alternative since it would have implied three times more subjects. The second possibility would have been to perform several randomly alternating blocks of each condition (with fewer items) instead of three main blocks assigned to each condition. We also discarded this second possibility in order to avoid production errors induced by the alternating conditions; however it remains a possibility worth to explore.

Second, we only have focused on token-to-token variability, and we selected articulatory data based on a single instant during vowel holding. Further insight may also be obtained by analyzing articulatory variability over the whole, sustained production of the vowel. Finally, our experimental study was designed in order to control and limit as much as possible additional sources of variability. However, this resulted in a highly constrained experimental design where subjects produce utterances in a strongly unnatural situation (bite block, EMA sensors, holding vowels). It is unclear how this impacts our results and analyzing production variability in more ecological conditions, for instance during sentence production, would certainly warrant further investigation.

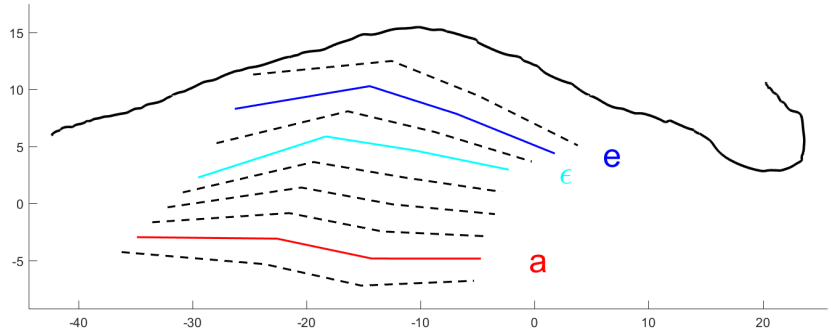

Figure 9.11: Illustration of target tongue postures for one subject (sagittal view). *The top black curve represents the subject's palate (front on the right). The colored lines correspond to vowel targets, identified to the corresponding average production of the subject during the production task. Black dashed lines correspond to the intermediate, non-vowel postures.*

3 Proprioceptive identification of vowels

3.1 Goals and constraints

The aim of this second study is to explore the accuracy of tongue proprioception for identifying vowel configurations. In order to address this question, the ideal situation would be to drive the tongue of participants towards different articulatory configurations and test (1) whether they are able to distinguish vowels from non-vowels configurations; and (2) in the case where the configuration corresponds to a vowel, whether they are able to correctly label it.

However, implementing this idea experimentally requires solving a crucial practical issue. Indeed, while driving an arm to different articulatory configurations may be relatively easy to perform with a robotic device, the situation is quite not the same for the tongue: the tongue is of difficult access inside the oral cavity, and, furthermore, it is a soft and slippery structure that is difficult to grasp and finely manipulate with a robotic device.

3.2 General description of the experimental design

To overcome this problem, we have considered the alternative of instructing subjects to drive their tongue themselves. To do so, we have designed a motor control task intended to guide subjects towards different tongue configurations while avoiding the involvement of speech related motor knowledge. The design of this motor control task intends to satisfy two main requirements: (1) it must lead subjects to reach a set of articulatory configurations among which some correspond to actual vowel configurations and some others do not; (2) it must avoid providing additional information that would enable the identification of tongue postures based on something else than proprioception.

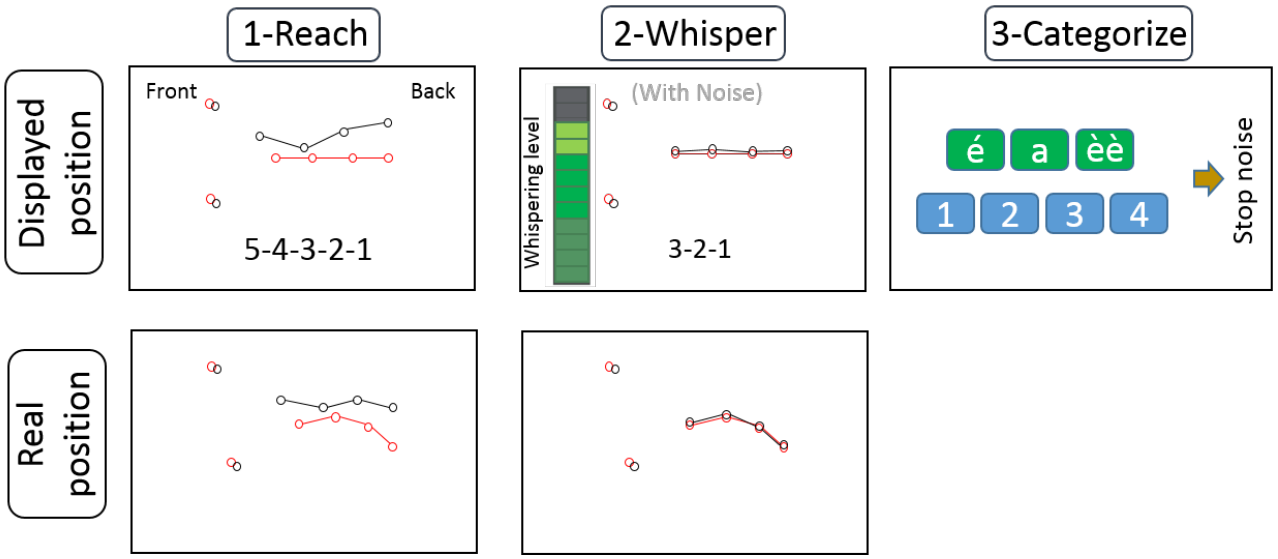

Figure 9.12: Trial design. *Each trial was composed of three tasks: (1) reaching a target tongue posture, (2) whispering the reached posture, and (3) categorizing the reached posture. Top panels illustrate the visual display presented to the subjects. Bottom panels illustrate the actual tongue configurations. Black circles correspond to real-time display of sensors positions, red circles correspond to the intended static targets. Vertically aligned circles on the left correspond to lip sensors.*

3.2.1 Definition of targets

In order to achieve the first requirement, we need to assess the specific tongue configurations corresponding to vowels for each subject. This was achieved with a preliminary vowel production task where tongue postures were recorded using the EMA system. This preliminary task was identical to the normal condition (N) of the production task in the previous study. Three vowels tongue postures (/e/, /ɛ/ and /a/) were kept from this preliminary production task. The choice of these vowels was motivated by the constraint to keep the reaching task simple. Each vowel tongue posture was estimated as the average tongue posture obtained from the 10 repetitions of each vowel. Tongue postures for each vocalization were selected automatically 0.7 seconds before the ending of the utterance. Six additional intermediate tongue postures were computed as equally spaced linear interpolations from the two closest vowel targets, out of the considered three (see Fig 9.11).

3.2.2 Reaching task

From this set of 9 target tongue configurations we designed a visually guided tongue reaching task. The task consisted of a visual display presenting the real-time position of the subject's tongue, measured with the EMA system, along with a static display corresponding to the target position to reach. Subjects were instructed to move their tongue in order to match at best its display with the target. In order to ensure that jaw position during the reaching task was the same as during the recorded targets, the overall task was performed while holding the same bite block as during the production task. Subjects were required to begin trials with their tongue in the /l/ position.

Concerning our second requirement, that of providing only somatosensory information to the subject during the task, the visual display was designed in order to avoid providing visual

information about tongue configuration, in terms of tongue position or shape. To do so, the visual display of both the target and subject's tongue were altered such that targets were always displayed as four horizontally aligned and equally spaced circles, each corresponding to one of the four EMA sensors on the subject's tongue (see Fig 9.12). This ensured that targets in all trial looked always the same, thus removing visual information about tongue shape and position. Furthermore, the position and movement of the subject's sensors (black circles in Fig 9.12) were displayed on the screen relative to the position of the corresponding target (red circles in Fig 9.12). In other words, the relative position of the black and red circles displayed on the screen was the same as the distance between the corresponding sensor and its target position in the real space. Consequently, since the red target circles were always displayed in the same horizontally aligned configuration, the resulting position of the black circles did not correspond to the actual physical configuration of the subject's tongue. Fig 9.12 illustrates the real tongue and target configurations and their corresponding visual display during the task. Subjects were instructed to move their tongue in order to match each black circle with the corresponding target red circle. When the match was achieved, the tongue of the subject reached the target configuration in real space.

During pilot tests, each trial of this reaching task would last as long as it would take to subjects for matching and holding target positions with enough precision. However, this appeared to be extremely difficult for most subjects, and most trials during pilot tests would take more than a minute to achieve. Therefore, in order make the task easier and limit the duration of each trial, subjects were given 5 seconds to get as close as they could get to the intended target. Remaining time was indicated with a countdown displayed on screen.

Note that the main goal of the task was to guide subjects to reach a range of different tongue postures among which certain are vowel postures and other intermediate tongue postures. Therefore, exactly reaching targets was not crucial. The main goal of the task was achieved as long as reached postures covered a uniform range of configurations that included the intended vowels.

3.2.3 Whispering task

A critical shortcoming of the EMA system is that it does not provide a complete account of the tongue. In particular, the back of the tongue is inaccessible since sensors cannot be fixed very far into the oral cavity. This raises the concern that the same positions of recorded sensors may be achieved by different tongue configurations. To address this issue, we further instructed subject to whisper while holding the reached posture. We expected that recording these whispered utterances would enable us to evaluate the correctness of the achieved postures. In order to prevent subjects from hearing their productions, their auditory feedback was masked during whispering, with an 80 dB(A) white noise sent through in-ear headphones. The masking noise began right after the end of the 5 seconds countdown of the previous reaching task. Subjects were instructed to whisper for around 1.5 seconds (duration displayed with a countdown on screen). In order to control the intensity of their whispering, subjects were provided with visual feedback through a level meter displayed on the screen (see Fig 9.12).

Lips were not constrained during the task. However, they may modify the produced sounds during whispering. In order to control this, the real-time positions of the lip sensors were also displayed, along with their corresponding target positions, recorded during the preliminary phase. Lip target positions were vertically aligned on the left of the screen and subjects were instructed to match them as best as they could during whispering (see Fig 9.12).

3.2.4 Proprioceptive identification task

For each trial, after the reaching and whispering tasks, subjects were instructed to label the identity of the reached posture. Answers were provided as a three alternative forced choice corresponding to target vowels /a/, /e/ and /ɛ/ (labeled respectively as "a", "é" and "èè"). Crucially, this identification task was performed right after the whispering task and subjects were instructed to answer based on the feeling of their tongue position during whispering. Subjects answered by clicking on the appropriate button on screen (see right panel of Fig 9.12), with the ordering of items displayed on buttons being randomized among trials.

Furthermore, in order to assess the accuracy of answers, subjects were instructed to rate their confidence about their answer based on a four-level scale (from 1, corresponding to "not confident at all", to 4 corresponding to "fully confident"). We expected that, if identification was accurate, subjects would answer with high confidence for postures that match vowel targets and with lower confidence for postures corresponding to intermediate postures. The masking noise was also maintained during this task in order to avoid that vowel identification would be influenced by eventual productions during this phase. The masking noise stopped right after subjects achieved this task.

3.2.5 Auditory identification task

Finally, after all trials of the proprioceptive task, subjects performed a final auditory identification task based on the whispered utterances they had produced. Subjects heard each of their produced whisper and were instructed to indicate which vowel best corresponded to the heard sound. Each whispered utterance was presented only once and their order was randomized. Answers for vowel labels and confidence levels were provided with the same three alternative forced choice design, and the same four-level scale as in the previous proprioceptive identification task.

3.3 Materials and Methods

The experimental protocol was performed in accordance with the ethical standards specified by the 1964 declaration of Helsinki and approved by the institutional ethics committee of the University Grenoble-Alpes.

3.3.1 Participants

The same 8 subjects as in the previous study participated in this study. All participants gave their informed consent before participating and were compensated with an additional 15€ gift cards.

3.3.2 Experimental session

The study took place in two consecutive days. On the first day, participants underwent a first training session to familiarize them with the reaching and whispering tasks. Crucially, the aim of this session was also to train subjects to perform the task without reference to any speech related knowledge. Hence, subjects were not aware of the aim of this training task, and no reference to speech was provided during this training session. This training session was performed right after the experimental session of the previous study and targets were defined based on productions performed during the normal condition (N) of this previous study. During this first training session, subjects were initially given 10 seconds, or more, in order to reach targets and this

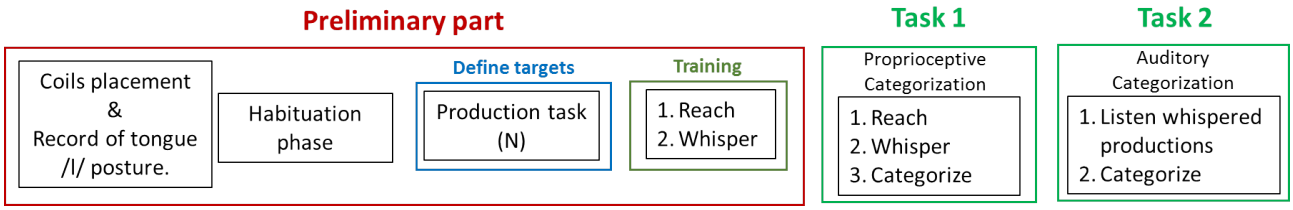

Figure 9.13: Block Diagram of the experimental design.

time was then progressively reduced to 5 seconds following subjects' improvements. In order to stimulate progression, an error score was attributed to each trial. Scores corresponded to the distance to the target and subjects were instructed to minimize this score. A bar plot indicating their improvement was displayed on screen.

The session on the second day was composed of two main parts (see Fig 9.13). **The preliminary part** was intended for recording target postures performed during a production task identical to the normal speaking condition (N) of the previous experiment. **The production task** was preceded by recording /l/ tongue configurations, as well as reading phonetically balanced questions, intended to accustom subjects with speaking with the sensors. The production task was followed by a short second training session intended to remind subjects of the reaching and whispering tasks on which they trained the day before.

After this short training, subjects performed the two main tasks of the study. Subjects began with the **proprioceptive categorization task**, where each of the 9 targets was displayed 10 times in a random sequence. The overall task thus contained 90 elementary trials composed of the reaching, whispering and identification sub-tasks.

Subjects ended the session with the **auditory categorization task**, performed after removing all sensors. Participants sat in a soundproof room and heard and labeled each of their 90 whispered productions presented through headphones (Sennheiser HD-25) in a randomized order.

3.3.3 Recordings and data processing

As in the previous experiment, articulatory data was collected with 4 midsagittally aligned EMA sensors (see Section 2 for details about sensor positioning), at 200 Hz and further filtered at 30 Hz for analysis. Articulatory configurations were kept for utterances during the production task and during the whispering portion of the reaching task in the same way as in the previous experiment: a representative tongue posture was selected manually by identifying a window of best articulatory stability and selecting the tongue posture at its center.

3.4 Preliminary results

3.4.1 Reaching task

The difficulty of the reaching task appears to be highly subject dependent. Some subjects had great difficulty in reaching target postures, while some others performed the task with ease. Figs 9.14 and 9.15 illustrate these differences. Fig 9.14 presents all the tongue postures achieved during the reaching task for two representative subjects, Subject 5 and Subject 8. Subject 5 had difficulties performing the reaching task, as can be observed from the larger variability of tongue postures as well as the greater distance between the reached and intended postures.

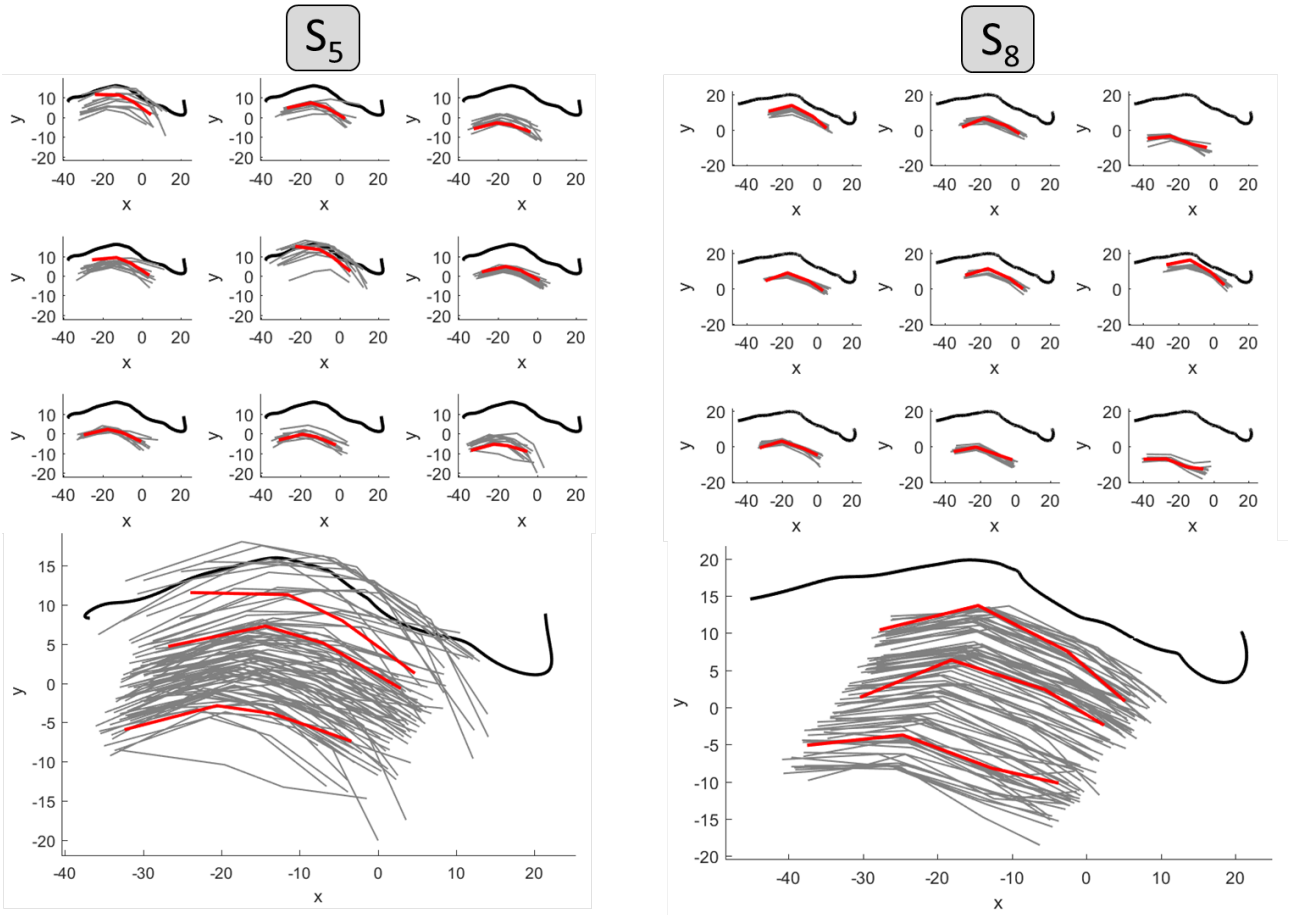

Figure 9.14: Illustration of reaching postures for Subject 5 (left) and Subject 8 (right). *Bottom panels present all reached postures (gray lines) with the three target vowel postures indicated for reference (red lines). Upper panels illustrate reached postures for each of the 9 intended targets. The first row corresponds to the three vowel targets, and the other two rows to the intermediate targets. The black curve on the top of each plot represents the sagittal trace of the palate (front on the right). The overlap between the tongue contours and the palatal trace for Subject 5 reveals that the palatal trace had not been measured exactly in the midsagittal plane of the subject (this was observed in some subjects, with stronger consequences for subjects having arched palates than for those having flat palates).*

Subject 8 performed the reaching task with ease, as can be observed from the regularity of reached postures and the better agreement between intended targets and reached postures.

Fig 9.15 presents a difficulty index quantifying the average performance of each subject during the reaching task. The difficulty index is computed as the average, across sensors, of their distances between reached and target positions. It can be seen that some subjects performed the task quite precisely (Subjects 4, 7 and 8) while others were less accurate. In particular, Subject 2 had great difficulties in reaching low tongue postures, hence the higher value of this subject's difficulty index.

3.4.2 Categorization task

Our aim is to assess the consistency of subjects' answers during the proprioceptive categorization task (called henceforth "proprioceptive answers", as opposed to "auditory answers" during

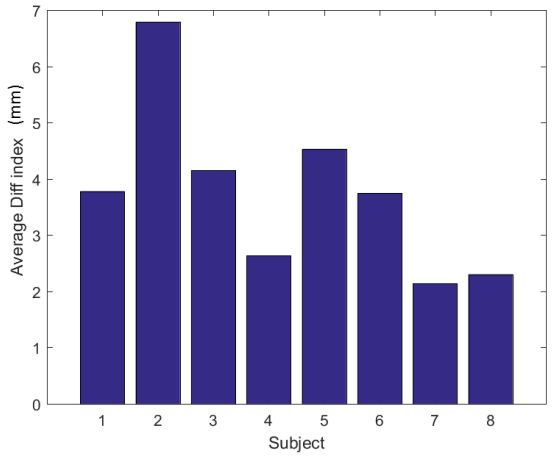

Figure 9.15: Difficulty index during the reaching task for each subject. *The score corresponds to the average across trials, sensors and targets, of the distance of each sensor to the corresponding target position.*

the auditory categorization task). A prerequisite for this is to determine how to evaluate consistency. In this section we present a number of approaches addressing this question. We begin by comparing proprioceptive answers with auditory answers and present results as confusion matrices summarizing the consistency of answers. However, a few studies of vowel perception in whispered speech have shown a significant decrease of the perception acuity, in particular along the /ɛ-a/ continuum (see for example Tartter, 1991), and when vowels are pronounced in isolation (Kallail & Emanuel, 1985). Consequently, the auditory answers of our subjects cannot be considered as an indisputable reference. Hence, it is important to have alternative evaluations that assess the consistency of proprioceptive answers with respect to their proximity to the considered vowel postures. We proceed in two steps: (1) a qualitative evaluation of this consistency by visualizing, for each subject, the auditory and proprioceptive answers associated with each tongue posture and observing the distribution of these answers within the geometrical domain covered by all tongue configurations measured for this subject; (2) a quantitative evaluation based, for each subject, on the comparison of the proprioceptive and auditory answers with the expected answers, as derived from a Bayesian model.

Comparing proprioceptive and auditory answers — Confusion matrices Fig 9.16 presents confusion matrices between proprioceptive and auditory categorization answers for each subject. The first salient observation is the strong variation among subjects of the percentage of matching answers. For S3 and S8, more than 85% of auditory and proprioceptive answers match. Chance level being at 33%, these scores are quite remarkable. Subjects 2, 4 and 5 have much lower matching scores (below 55%). Another noteworthy observation is the differences between percentages of matching answers for each vowel. For many subjects, there is a bias toward a more frequent perception of vowel /a/ and a less frequent perception of vowel /ɛ/. In addition, the consistency between proprioceptive and auditory identification of vowel /ɛ/ is weak.

We also evaluated the potential relation between motor dexterity, as described by the difficulty index, and the consistency of auditory and proprioceptive answers. Fig 9.17 presents the corresponding results. It can be seen that there is no clear relation between the difficulty index and the matching score. The slight negative correlation of the linear regression appears

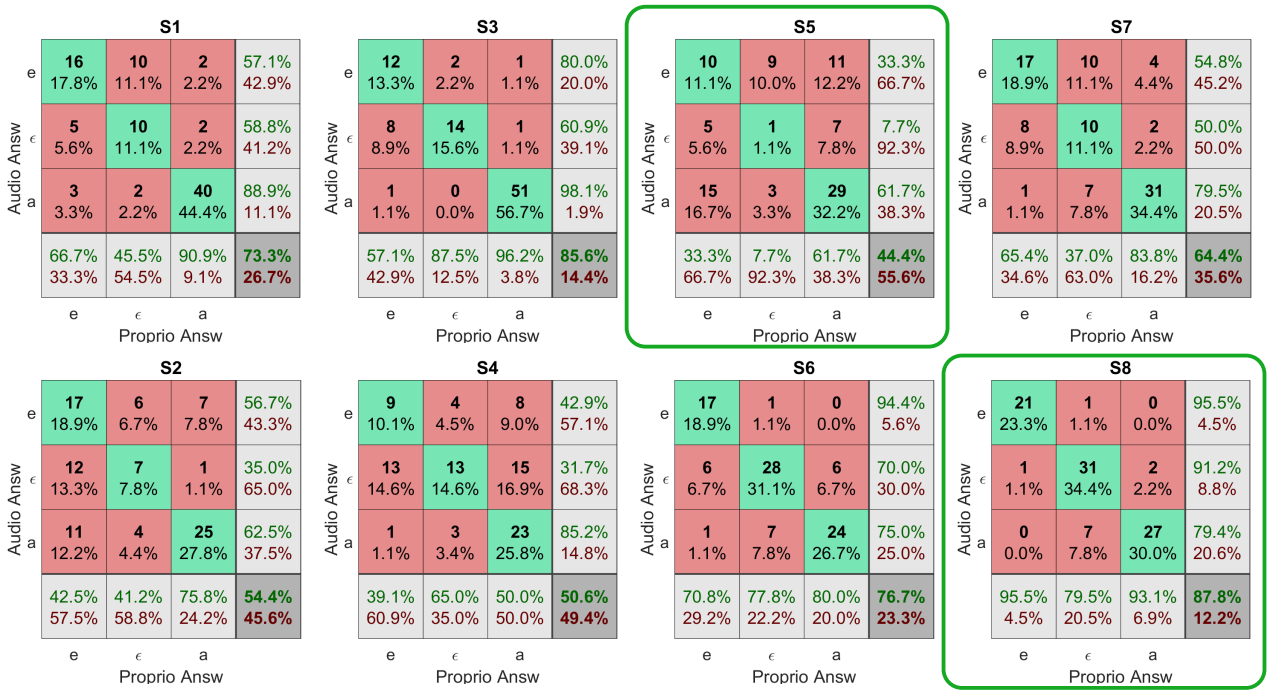

Figure 9.16: Confusion matrices comparing proprioceptive and auditory categorization answers for each of the 8 subjects. *In the confusion matrix (the 3 upper lines and 3 most left columns), bold characters indicate the number of stimuli that have been categorized within each combination of the auditory and somatosensory categories. Percentage values written in black indicate the proportion of the total number of trials (90) that has been categorized within each of the 9 combinations of auditory and somatosensory categories. Outside the boundaries of the confusion matrix, percentages written in green represent the percentage of trials in the line (resp. the column) that have been categorized similarly in both conditions. The percentages written in red represent the percentage of trials that have been categorized differently. Subjects S5 and S8 (highlighted by green frames) are illustrative examples of two extreme cases among the eight subjects.*

not to be significant (p=0.432), a result corroborated by non significant Pearson and Spearman correlation coefficients (p=0.432 and p=0.428 respectively).

Qualitative evaluation of the consistency of auditory and proprioceptive answers
An alternative approach consists in evaluating for each subject the consistency of the categorization answers in relation to the tongue configuration underlying the production of the vowel stimulus. Fig 9.18 presents, for subjects S5 and S8, the total set of reached postures along with the subsets of tongue configurations attributed to each vowel in the proprioceptive (top panels) and auditory (middle panels) categorization tasks. Variability across subjects is clearly exemplified: for S8, tongue postures assigned to each vowel are neatly separated, whereas for S5 they largely overlap in both modalities.

While the visualization of tongue postures for each answer gives a qualitative idea of the consistency of answers, it is difficult to really assess the proximity of tongue postures to the reference vowel postures. Indeed, tongue postures are measured by 4 sensors, which provide 4 distance measures relative to reference positions. An alternative approach is to represent tongue postures in the principal component space computed from all vowel postures (see e.g., Harsh-

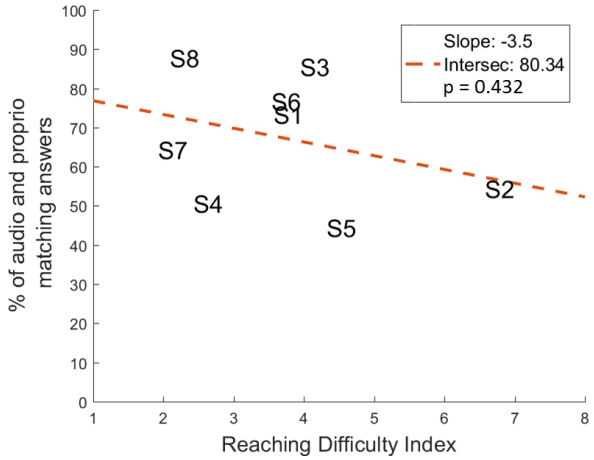

Figure 9.17: Percentage of matching auditory and proprioceptive answers as a function of the reaching difficulty index for each subject.

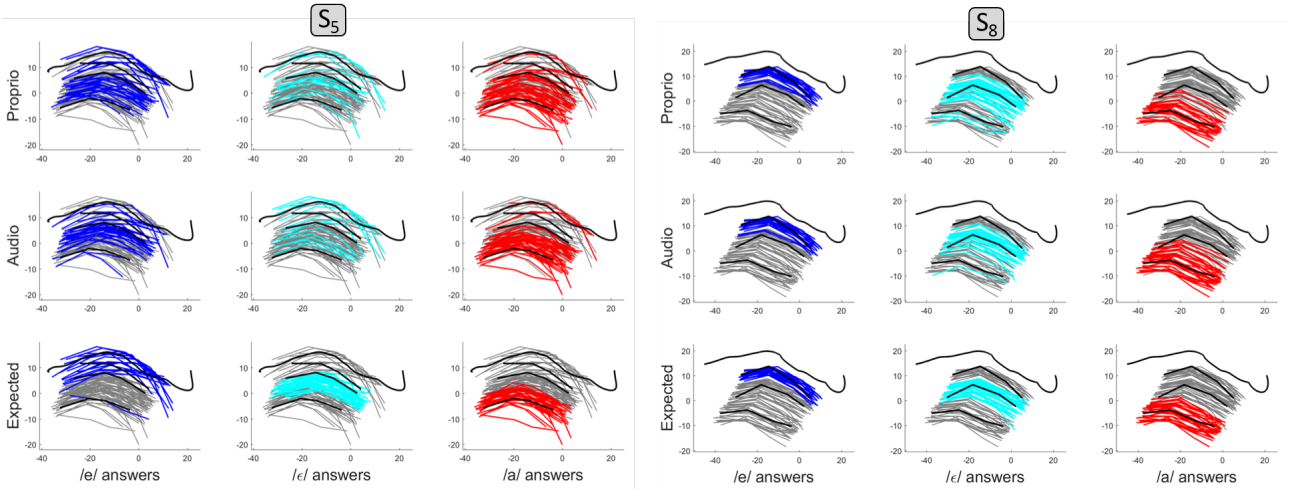

Figure 9.18: Illustration of categorization results for Subject 5 (3 most left columns) and Subject 8 (3 most right columns). *Top row: Proprioceptive answers; Middle row: Auditory answers; Bottom row: Expected answers (see below). Gray background lines correspond to the total set of reached postures. Colored lines correspond to postures labeled according to the vowel indicated in each column.*

man, Ladefoged, & Goldstein, 1977; Hoole, 1998; Jackson, 1988; Maeda, 1990; Nix, Papcun, Hogden, & Zlokarnik, 1996; Whalen, Chen, Tiede, & Nam, 2018). This enables to represent tongue configurations with respect to the two most salient "shape directions".

Fig 9.19 presents the same data as in Fig 9.18, but projected onto the subject-specific 2-dimensional PCA space, with dispersion ellipses (± 1 standard deviation) corresponding to the productions of each vowel during the production task. As in Fig 9.18, it can be seen that S8 consistently associated tongue configurations and auditory or somatosensory categories, whereas answers of S5 are overlapping and less consistent with the expected vowel regions. Interestingly, in line with the confusion matrices shown above, there is no clear difference in categorization according to whether auditory feedback was available or not.

The degree of confidence attributed by subjects to their answers is given in Fig 9.18 by the

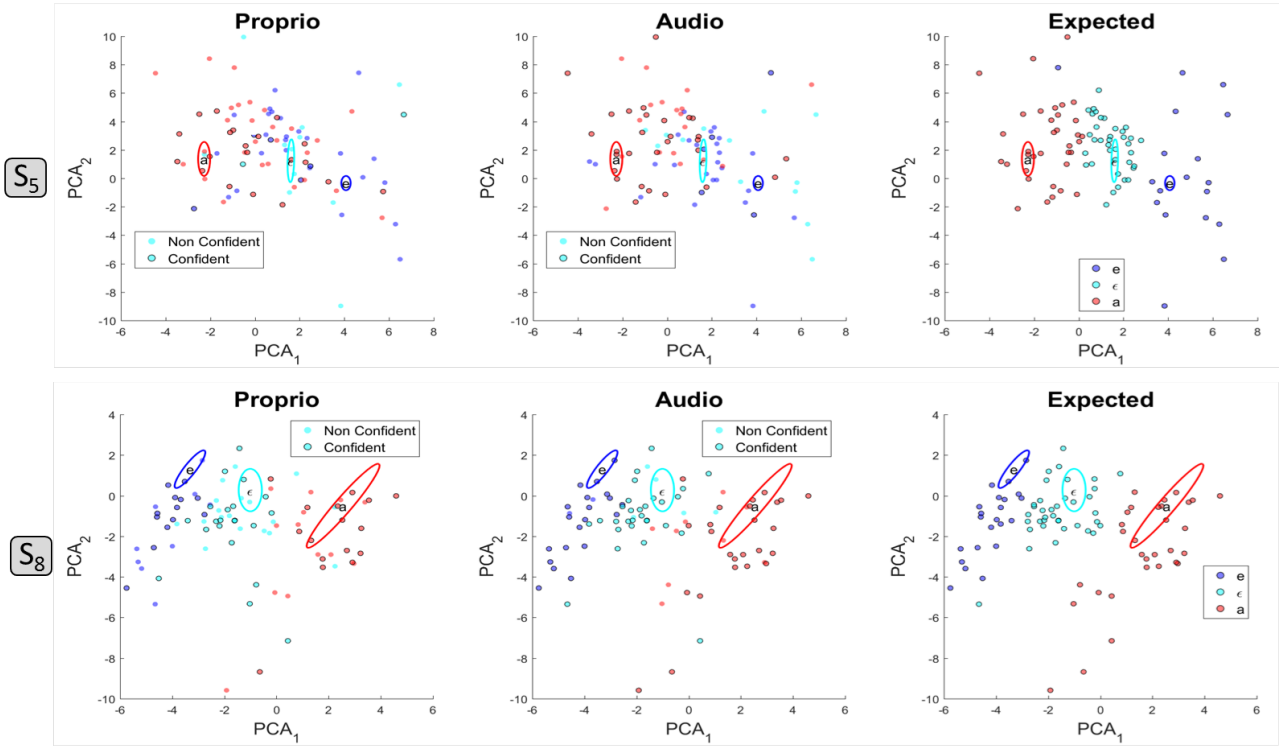

Figure 9.19: Illustration of auditory and proprioceptice answers in the subject specific PCA space for Subject 5 (top panels) and Subject 8 (bottom panels). *Left column: Proprioceptive answers; Middle column: Auditory answers; Right column: Expected answers (see below). Blue circles correspond to vowel /e/, cyan circles to vowel /ɛ/, and red circles to vowel /a/. Confident answers are represented by markers with black colored edges.*

type of marker used: markers with a black colored edge correspond to high levels of confidence. It can be seen that, surprisingly, there does not appear to be any obvious relation between the proximity of the tongue shape underlying the stimulus with the tongue shape measured for each vowel and the degree of confidence attributed to the answer. We would have expected high degrees of confidence to be associated with tongue configurations inside dispersion ellipses.

Quantitative evaluation with respect to the expected answers

Bayesian model Representing the data in the speaker-specific PCA spaces, as in Fig 9.19, suggests a theoretical approach to characterize the probability of assigning a given label to each tongue posture. More specifically, we define a Bayesian classifier that provides the probability $P(\Phi \mid S)$ that a given tongue shape S is identified as phoneme Φ. The term $P(\Phi \mid S)$ can be defined as an inference process obtained from a model based on the distribution of the tongue postures produced for each vowel during the production task. Indeed, each elliptic region in Fig 9.19 enables to define terms $P(S \mid \Phi)$ as Gaussian probability distributions with mean and covariance matrices corresponding to each region. From these terms, and assuming uniform probabilities for $P(\Phi)$, we compute[6]:

$$P([\Phi = \phi] \mid [S = s]) \quad = \quad \frac{P([S = s] \mid [\Phi = \phi])}{\sum_{\phi'} P([S = s] \mid [\Phi = \phi'])}. \tag{9.1}$$

[6]This is essentially the same as the categorization model defined in Chapter 5, but in somatosensory rather than auditory space.

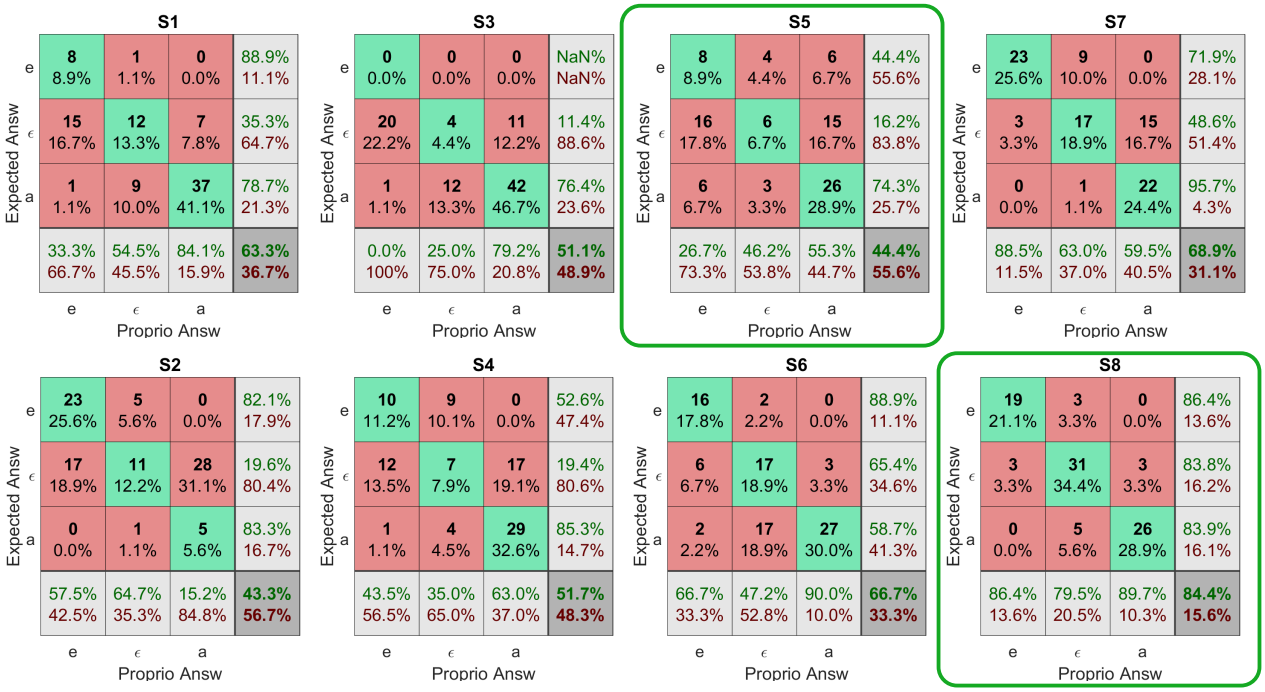

Figure 9.20: Confusion matrices comparing proprioceptive answers with respect to expected answers. *See Fig 9.16 for further information.*

Visualization of expected answers The Bayesian classifier enables to define a decision procedure, in order to attribute to each tongue posture an expected vowel answer. This is done by attributing to each tongue posture s the label ϕ that has highest probability in $P([\Phi = \phi] \mid [S = s])$. The bottom row of Fig 9.18 and the right column of Fig 9.19 present the expected vowel answer for each tongue posture, for Subjects S5 and S8.

Comparison with expected answers We can now compare auditory and proprioceptive answers with respect to the answers predicted by the Bayesian classifier. Fig 9.20 presents confusion matrices for each subject between proprioceptive and expected answers, and Fig 9.21 presents confusion matrices for each subject between auditory and expected answers. The total matching scores indicated in Figs 9.16, 9.20 and 9.21 are further summarized in the bar plot of Fig 9.22.

Figures 9.20 and 9.21 confirm the trends provided previously by the confusion matrices between proprioceptive and auditory answers (Fig 9.16): (1) globally, categorization based on proprioceptive feedback only is close to auditory categorization in whispered speech; this is true for all the subjects except S2; (2) a bias toward the /a/ category exists for the majority of subjects both for proprioceptive and auditory answers. In addition we note that, in the auditory identification task, the majority of subjects tend to identify /ɛ/ or /a/ rather than /e/.

Figure 9.22 shows that two subjects clearly stand out from these global observations: S2 and S3. While S2 provides proprioceptive answers that are in line with the rest of subjects, S2 shows in the auditory identification task a very strong bias to the intermediate /ɛ/ category (>60% of the answers). The case of S3 is particular: her global matching score between auditory and proprioceptive answers is the second best (85.6%), while her matching scores with respect to expected answers are much lower (around 50%). This is explained by the fact that no /e/ label

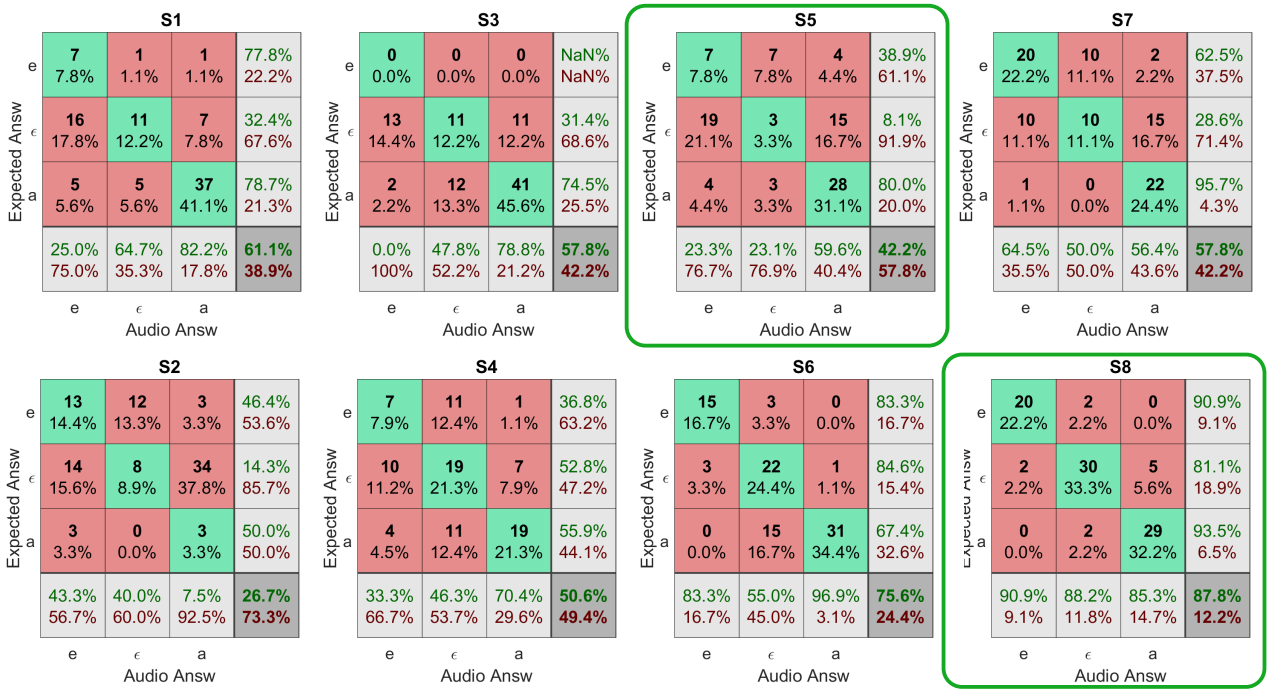

Figure 9.21: Confusion matrices comparing auditory answers with respect to expected answers. *See Fig 9.16 for further information.*

is predicted by the Bayesian classifier, as can be seen in Figs 9.20 and 9.21. This is due to a small variability of the productions of vowel /e/ for this subject, and thus the inferred Gaussian distribution describing the dispersion ellipsis is narrow, and none of the reached tongue postures is sufficiently close to this theoretical ellipsis to be classified as belonging to this distribution.

Fig 9.23 further relates the total matching score in each condition with the difficulty indexes

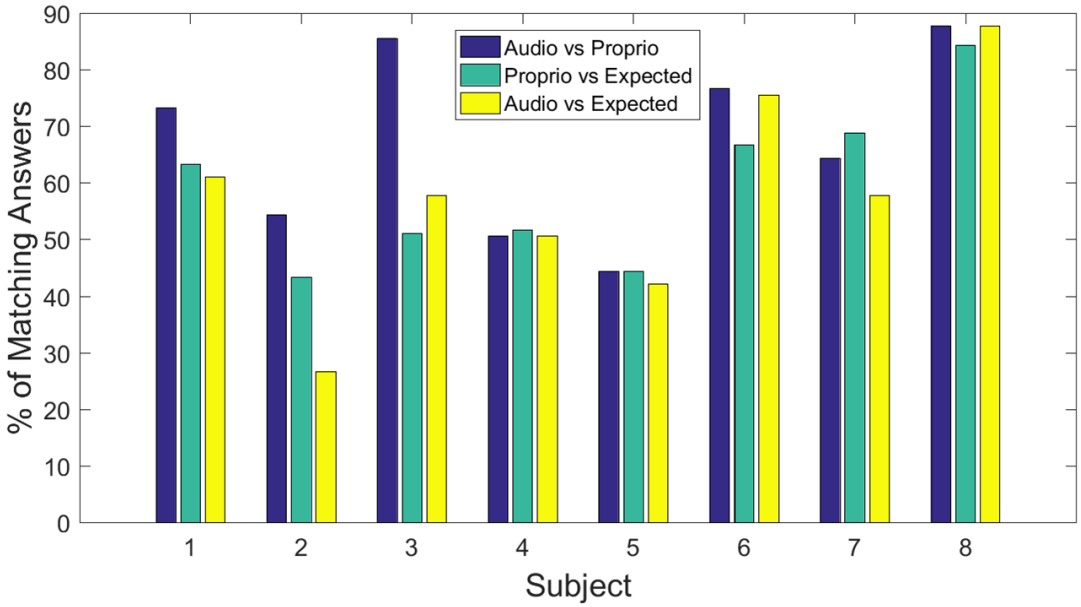

Figure 9.22: Comparison of matching scores for each subject.

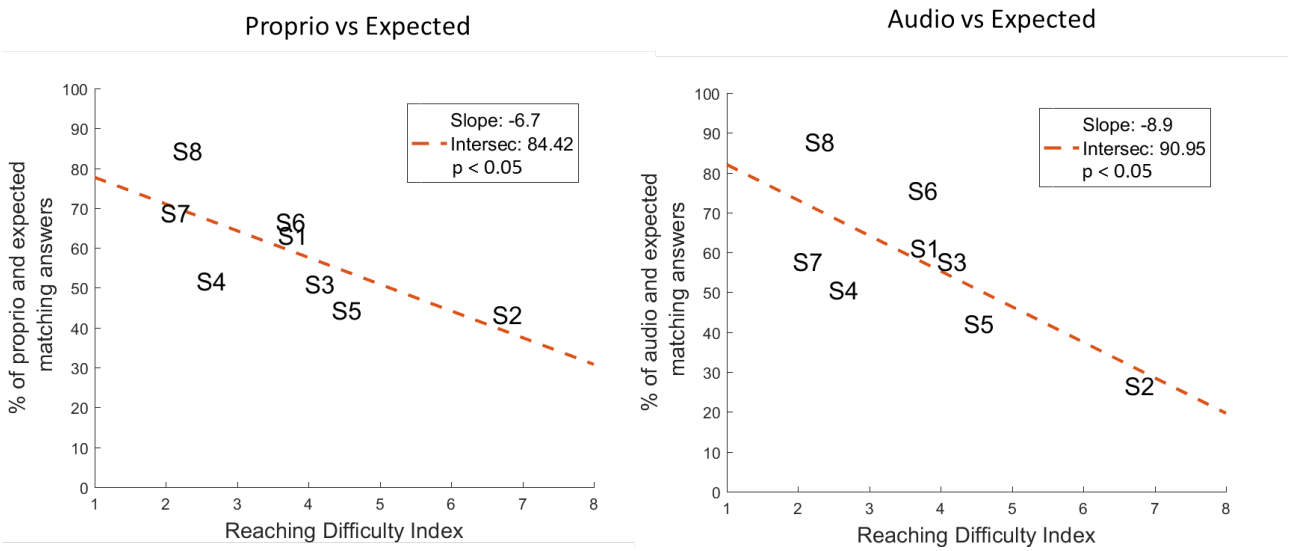

Figure 9.23: Correlation between percentage of matching answers and reaching difficulty index. *Left: proprioceptive vs expected answers; Right: auditory vs expected answers.*

presented in Fig 9.15. It can be seen that there is a significant negative correlation between difficulty indexes and matching scores in both auditory and proprioceptive cases.

Overall, it is interesting to note that the global identification rates reached by the subjects in the auditory identification task are in the range of the identification rates found by Kallail and Emanuel (1985) for isolated whispered vowels in American English, except, again, for subject S2.

Visualizing categorization curves The general objective of this experiment is to assess the capacity for subjects to categorize tongue configurations into phonemes along the /e-ɛ-a/ continuum based on proprioceptive feedback only, and to compare this capacity with the well-known categorical perception in the auditory domain. To do so, we use procedures that are similar to the ones used in perceptual categorization tests and enabling to draw categorization curves.

In order to describe categorization curves along a single dimension, in agreement with our Bayesian analysis, we decided to display categorization as a function of the Mahalanobis distance separating the tongue posture from the center of one arbitrarily chosen vowel category. The Mahalanobis distance quantifies the distance from the mean of a distribution, scaled by the standard deviation of this distribution. With this distance "relative to variance", we can define elliptic rings that provide a topography of the domain in relation to the considered vowel region, as illustrated for the /e/ region in Fig 9.24. Expected, proprioceptive and auditory categorization curves can then be obtained as a function of the Mahalanobis distance by computing the proportion of, respectively, expected, proprioceptive, and auditory answers within two consecutive rings. Fig 9.25 illustrates these three categorization curves for subjects S5 and S8.For subject S8, the match between both auditory and proprioceptive answers with the expected, theoretical answers is striking. In contrast, categorization curves for subject S5 are much less clear, even though there is a small trend for answers /a/ and /e/ to be consistent with their corresponding expected answers.

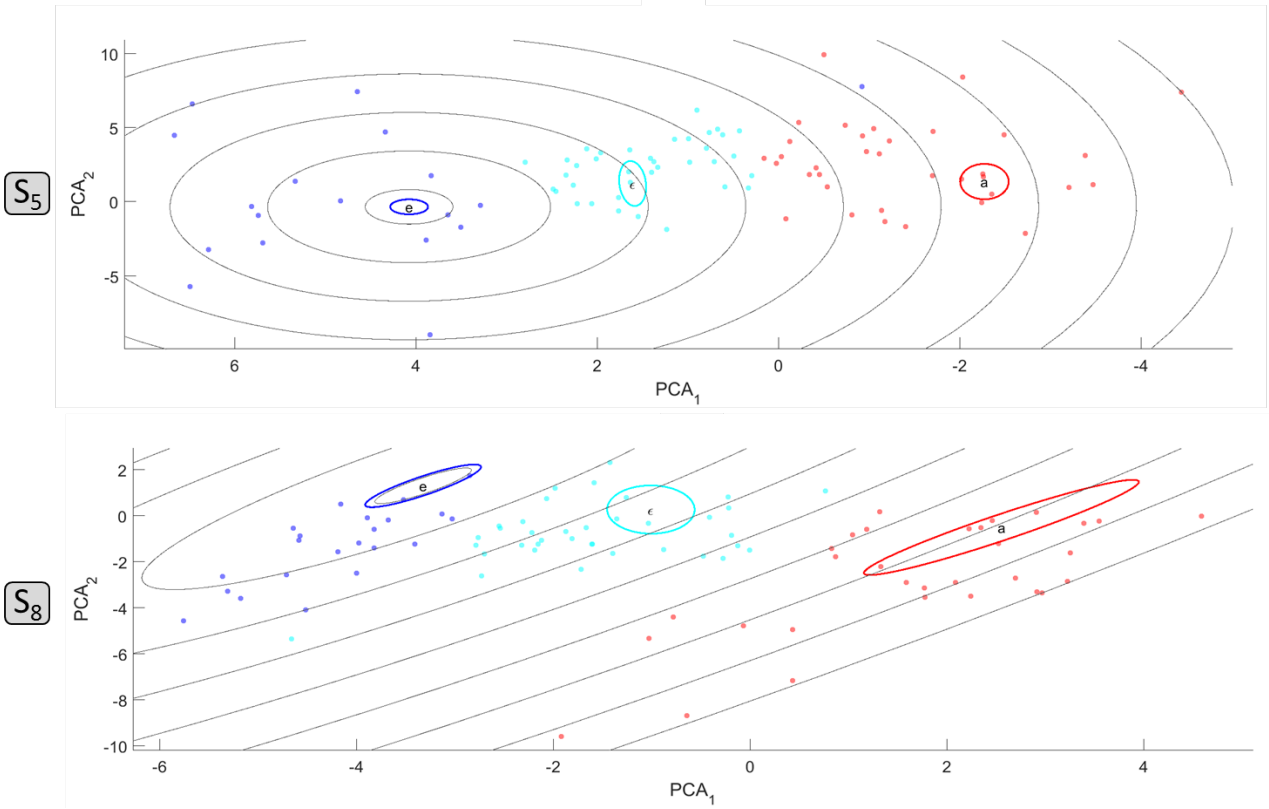

Figure 9.24: Illustration of the rings providing a topography of the domain in relation to the /e/ category for subjects S5 and S8. *Regions are defined by 9 equally spaced rings between the closest and furthest data points in relation to the /e/ category. Experimental data are colored according to the expected answers, as determined by the Bayesian model.*

Evaluating the consistency of subjects categorization with model comparison Finally, Bayesian models enable to evaluate the consistency of subject's categorization by comparing the relative log-likelihood of their answers with respect to different models. The first model that we consider corresponds to the situation where subjects would be unable to identify vowels and would thus categorize at random, with each label being equally probable. We implement this idea with a uniform model that attributes probability 1/3 to each label for every tongue configuration. We compare this uniform model to two versions of the categorization process defined in Eq (9.1). These versions implement the hypothesis that categorization would be precise and consistent with the Bayesian classifier for tongue postures that are close to the production regions, whereas, for tongue postures that are far from the productions regions, subjects would not be able to identify them and thus would answer at random with uniform probability 1/3 for each label. The two versions are obtained by modifying the Gaussian probability distributions $P(S \mid \Phi)$ in the following way: values of $P(S \mid \Phi)$ that are smaller than a given threshold are set to the threshold value, while other remain unchanged. The value of the threshold for the two considered versions was set to the probability value at one and two standard deviations respectively.

Fig 9.26 presents – for proprioceptive, auditory and expected answers – the three relative log-likelihoods comparing the three considered models, in three two-by-two comparisons. The sign of relative log-likelihoods indicates which model in the comparison M1-M2 better accounts for experimental observations; positive values indicate that model M1 better agrees with data than model M2. Comparisons in the case of expected answers (i.e., answers labeled according to the

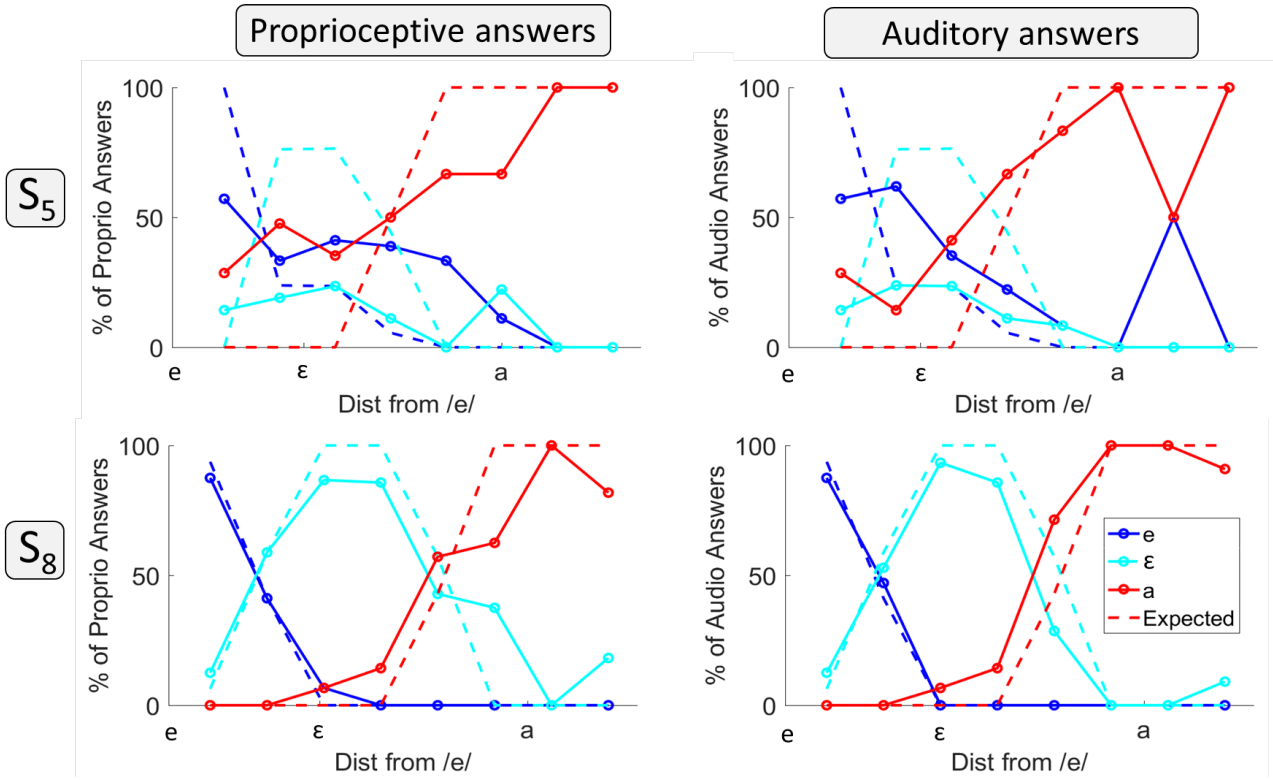

Figure 9.25: Categorization curves as functions of the Mahalanobis distance from vowel /e/. *Left panels: proprioceptive and expected categorizations; Right panels: auditory and expected categorizations.*

theoretical Bayesian classifier) are provided for reference. As expected, comparing models with respect to answers predicted by the Bayesian classifier indicates that the two modified Bayesian classifiers are better descriptors than the model based on uniform distributions (positive yellow and green bars on the right panel of Fig 9.26), with the Bayesian classifier modified at one standard deviation being closer to the uniform model (smaller green bars than yellow bars), as expected from its definition. Also, the Bayesian classifier modified at two standard deviations better agrees with the expected answers than the Bayesian classifier modified at one standard deviation (negative blue bars in the comparison 1std-2std on the right panel of Fig 9.26), as expected from their definition.

Comparing models based on subject data provides a measure of how consistent their answers are with respect to the theoretical Bayesian classifier. The positive green and yellow bars on the left and middle panels of Fig 9.26 indicate that, for most subjects, the Bayesian classifiers modified at one and two standard deviations are clearly better descriptors of their answers than the model based on uniform distributions, for somatosensory and to a lesser extent for auditory answers. The large negative differences between the classifier modified at one standard deviation and the one at two standard deviations (blue bars on the left and middle panels of Fig 9.26) indicate that consistency of subject answers with respect to the Bayesian classifier is not limited to a close proximity of vowel regions, but extends to tongue postures in the range between one and two standard deviations away from average productions.

Overall, this model comparison approach suggests that our experimental data can be accounted for, and are consistent with, a model of somatosensory perceptual categorization based on Gaussian somatosensory prototypes, paving the way towards confirming that somatosensory

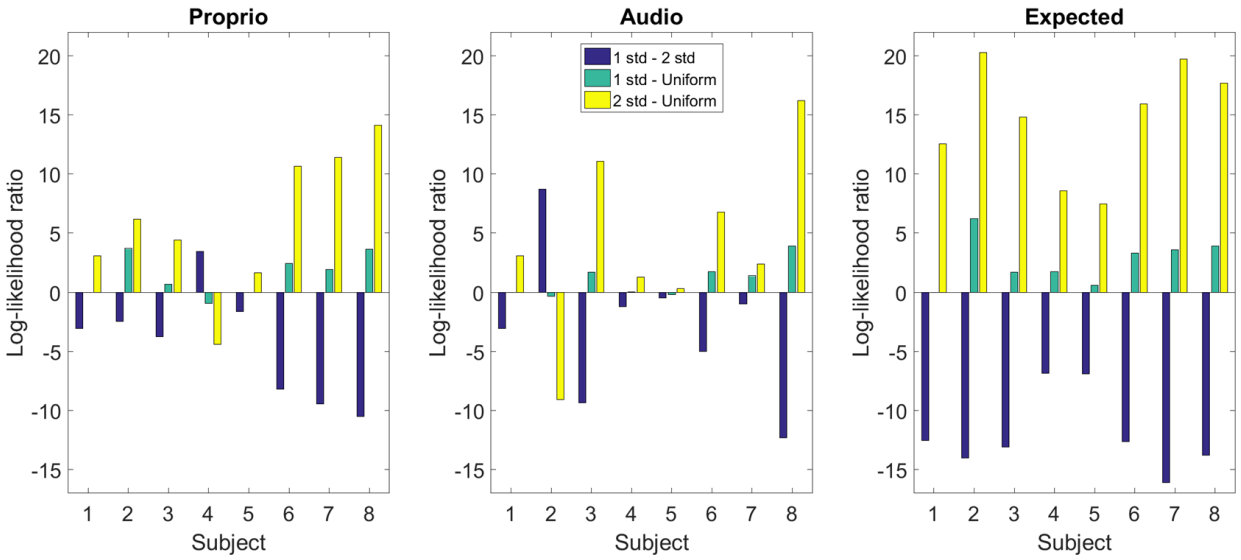

Figure 9.26: Relative log-likelihoods between considered models for each subject based on proprioceptive, auditory and expected answers. *Blue bars: comparison of Bayesian classifier modified at one standard deviation versus Bayesian classifier modified at two standard deviation; Green bars: comparison of Bayesian classifier modified at one standard deviation versus model based on uniform distributions; Yellow bars: comparison of Bayesian classifier modified at two standard deviations versus model based on uniform distributions.*

perception would be categorical.

3.5 Discussion

3.5.1 Summary of results

We have designed an experimental study in order to assess the existence and precision of proprioceptive characterizations of vowels. The protocol includes a reaching task followed by a categorization task were subjects were instructed to identify the reached tongue posture. Results of the reaching task indicate that there are strong individual differences in the ability to finely control the tongue, when guided only by visual feedback. The difficulty of visually guided tongue control is well known, in particular in the field of computer assisted language learning (Fabre, 2016; Ouni, 2014).

We evaluated the consistency of proprioceptive categorization with qualitative and quantitative approaches. These analyses indicate two main results. Firstly, we observed important inter-subject differences in the consistency of proprioceptive categorization answers with respect to expected answers: some subjects performed highly consistently and in agreement with the expected answers, while others performed poorly, with a less regular pattern of answers and a small agreement with expected answers. Secondly, despite these important inter-subject differences, the overall pattern of proprioceptive and auditory answers were similar in each subject: subjects that performed consistently in the proprioceptive categorization task also performed consistently in the auditory categorization task in whispered speech, and *vice versa*.

Overall, our results suggest that the accuracy of perceptual categorization based on somatosensory feedback is similar to the accuracy of auditory perception of whispered vowels. Since categorical auditory perception has been shown to be significantly worse with whispered

speech than with modal speech (e.g., Tartter, 1991), our experiment suggests that the perception of one's own speech production based on somatosensory feedback is weakly categorical and less efficient than in auditory processing of modal speech.

We further evaluated a possible relation between the dexterity of tongue control and the accuracy of proprioceptive categorization, where dexterity was assessed in terms of a difficulty index in performing the reaching task, and accuracy of proprioceptive categorization was assessed in terms of the overall consistency of answers with respect to expected answers. We observed a weak negative correlation between reaching difficulty and accuracy of proprioceptive categorization. While further subjects would be certainly needed in order to better specify this trend, it is important to note that this negative correlation does not imply a relation between the dexterity of tongue control and the accuracy of proprioceptive categorization. Indeed, our results also indicate a similar negative correlation between reaching difficulty and accuracy of auditory categorization. This could suggest that subjects with poor tongue dexterity may have worse categorization capabilities, both proprioceptive and auditory (in isolated whispered utterances). However, an alternative and more plausible explanation is that subjects that struggled during the reaching task performed a greater number of awkward tongue postures, resulting in whispered utterances that were more difficult to categorize. In order to test this hypothesis, one option would be to present the same whispered stimuli to other subjects and assess whether the low categorization scores we observed are due to intrisically difficult stimuli, or to our subject's poor categorization abilities.

An additional question in the interpretation of our results is to assess to what extent proprioceptive categorization is based on proprioceptive targets or on an internal auditory simulation based on proprioceptive information. We expected that the presence of noise during the proprioceptive categorization task would prevent subjects from hearing their whispers as well as limiting their access to internal auditory simulations. However, we cannot preclude this case, and an interpretation based on auditory simulation during categorization cannot be discarded. However, further work could be developed in this direction. For instance, we could exploit non linearities in the relation between tongue postures and acoustic outputs. These non linearities may result in different categorical boundaries between vowels in articulatory and auditory spaces. If this is indeed the case, it would provide evidence for identification based on two different sensory targets.

3.5.2 Caveats and future directions

While our results are certainly promising, a certain number of points would deserve to be further explored. We begin by summarizing additional analyses of our current data that may enable to gain further insight on our results. Then we consider possible improvements of the protocol and future directions to explore.

Further analyses The first point that certainly needs to be further analyzed is the reliability of tongue postures used in our analyses. Tongue postures during the reaching task were selected based on a single instant during whispering. Some subjects were able to maintain their posture with greater stability than others. A more detailed analysis of tongue posture variability during whispering would be important in order to assess the reliability of our measure. This analysis remains to be done.

A second point that remains to be analyzed is the stability of lip position during the whispering task. We provided visual feedback of lip sensors in order to help subjects keep their lips stable. We still need to confirm that subjects actually maintained their lips stable.

Finally, we have assessed the difficulty of the reaching task with an index measuring the distance between the reached posture and the intended target. It would be important to find other indices to improve this measure. Analysis of tongue trajectories towards the intended targets may provide some interesting clues.

Possible improvements One of the main difficulties of our experimental protocol is the limited information that is accessible with EMA sensors. In particular, the transverse configuration of the tongue along its most posterior part is inaccessible. This prevents us from precisely assessing the real proximity of performed tongue postures with respect to target postures. While this shortcoming may have been avoided by using an echograph, it would have raised additional technical difficulties in the real-time extraction of tongue contour in order to provide real-time visual feedback to the subject. However, the method could have been used in addition to the EMA sensors in order to complete the missing information of tongue shape during analyses. We did not include this in our protocol for simplicity, but it could be a possibility to be considered in the future.

Following this same concern, note that we evaluated the consistency of answers only with respect to those expected by articulatory similarity with tongue postures during production. It would have been interesting to perform the same evaluation but with respect to the spectral characteristics of the performed whispers, comparing them with the spectral characteristics of vowel productions. However, whispered speech is noisy and difficult to analyze. We explored the possibility of estimating these spectral characteristics indirectly with an impedencemeter (Epps, Smith, & Wolfe, 1997), however we did not pursue this idea further due to the additional load to the experimental protocol. It also remains an idea worth to be further explored.

We should also evaluate to what extent categorization after the reaching task involved only somatosensory information. Our experimental protocol intended to visually guide subjects to drive their tongue toward different tongue configurations. Crucially, our main concern was to avoid providing visual information concerning tongue shape and position, and to do this, we altered the visual display of targets and tongue shape such that all targets looked the same on screen. However, a possible weakness of our protocol is that the visual display still contained information about the relative position of the target with respect to the tongue. Since in our protocol we instructed subjects to begin their reaching movement always from the same starting position (the /l/ tongue configuration), subjects may have used this relative visual information in order to identify targets on the basis of the visual path from the onset to the target. While it would be important to evaluate more precisely this feature in future developments of the protocol, we believe that this did not dramatically influence our results. Indeed, subjects were asked to report their strategy during the reaching and categorization tasks, and only one subject (S6) reported to have thought, at some point, that the relative visual position in the screen may have helped to identify /e/ items, due to their proximity with the initial /l/ configuration. In order to avoid this eventual shortcoming, one possibility would be to lead subjects to different starting positions prior to each trial of the reaching task.

Future directions In the present study we chose to focus on a limited number of vowel targets in order to simplify the task and limit the duration of experimental sessions. The choice of vowels /e-ɛ-a/ was motivated by two main reasons. Firstly, we focused on an articulatory continuum that could be thought of as 1-dimensional. The /e-ɛ-a/ continuum, in this regard, seemed to be the most natural choice. Secondly, during preliminary tests, pilot subjects seemed to have trouble reaching the tongue configurations of /ɔ/ and /œ/, which are more posterior. Hence, we decided to keep targets /e-ɛ-a/ to make the reaching task easier. However, it remains

unclear why some tongue configurations were more difficult to reach than others, and why some subjects struggled more than others. It would be interesting to further develop the reaching paradigm in order to explore these questions.

A final question that could also be worth exploring with our reaching paradigm is the precision of tongue proprioceptive acuity. We informally asked the three last subjects how many different tongue targets they thought they had to reach. Interestingly, subjects tended to underestimate the number of targets. Whereas all subjects had 9 different targets to reach, subject S6 reported 5 targets, subject S7 reported 3 targets and subject S8 initially reported 5 and later 8. It is unclear to what extent these answers may have been influenced by the three forced choice items of the categorization task. In order to avoid this issue, this could be assessed during the preliminary training session, when no reference to categorization is given. In particular, the reaching task may be used in order to estimate the minimum distance between targets that subjects are able to distinguish, and if this distance is the same all across vocal space, or if it varies with respect to some tongue configurations. Such a discriminability analysis could complement our current study to better assess categorical perception of somatosensory information.

Part III

Perceptuo-motor interactions in speech

Chapter 10

What drives the perceptual change following speech motor adaptation? Evaluation of hypotheses in a Bayesian modeling framework

Introduction

This chapter presents our third contribution, in which we examine perceptuo-motor interactions in speech. This contribution, as a whole, was the object of a recent publication (Patri et al., 2018). The model it features is a simplified one-dimensional implementation of the multisensory Bayesian model we have developed in previous chapters. In terms of model definition therefore, there is not much overlap between the material presented previously and the one described in this chapter. For convenience, we have thus opted to include the published paper, in extenso, in the following pages. The model is used to evaluate different hypotheses that may account for recent findings on perceptual shifts resulting from speech motor adaptation.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

What drives the perceptual change resulting from speech motor adaptation? Evaluation of hypotheses in a Bayesian modeling framework

Jean-François Patri[1,2]*, Pascal Perrier[1], Jean-Luc Schwartz[1], Julien Diard[2]

1 Univ. Grenoble Alpes, CNRS, GIPSA-Lab UMR 5216, F-38000 Grenoble, France, **2** Univ. Grenoble Alpes, CNRS, LPNC UMR 5105, F-38000 Grenoble, France

* jeanfrancoispatri@gmail.com

Citation: Patri J-F, Perrier P, Schwartz J-L, Diard J (2018) What drives the perceptual change resulting from speech motor adaptation? Evaluation of hypotheses in a Bayesian modeling framework. PLoS Comput Biol 14(1): e1005942. https://doi.org/10.1371/journal.pcbi.1005942

Editor: Frédéric E. Theunissen, University of California at Berkeley, UNITED STATES

Received: July 25, 2017

Accepted: December 26, 2017

Published: January 22, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no. 339152, "Speech Unit(e)s", PI: Jean-Luc-Schwartz). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Shifts in perceptual boundaries resulting from speech motor learning induced by perturbations of the auditory feedback were taken as evidence for the involvement of motor functions in auditory speech perception. Beyond this general statement, the precise mechanisms underlying this involvement are not yet fully understood. In this paper we propose a quantitative evaluation of some hypotheses concerning the motor and auditory updates that could result from motor learning, in the context of various assumptions about the roles of the auditory and somatosensory pathways in speech perception. This analysis was made possible thanks to the use of a Bayesian model that implements these hypotheses by expressing the relationships between speech production and speech perception in a joint probability distribution. The evaluation focuses on how the hypotheses can (1) predict the location of perceptual boundary shifts once the perturbation has been removed, (2) account for the magnitude of the compensation in presence of the perturbation, and (3) describe the correlation between these two behavioral characteristics. Experimental findings about changes in speech perception following adaptation to auditory feedback perturbations serve as reference. Simulations suggest that they are compatible with a framework in which motor adaptation updates both the auditory-motor internal model and the auditory characterization of the perturbed phoneme, and where perception involves both auditory and somatosensory pathways.

Author summary

Experimental evidence suggest that motor learning influences categories in speech perception. These observations are consistent with studies of arm motor control showing that motor learning alters the perception of the arm location in the space, and that these perceptual changes are associated with increased connectivity between regions of the motor cortex. Still, the interpretation of experimental findings is severely handicapped by a lack of precise hypotheses about underlying mechanisms. We reanalyze the results of the most

Competing interests: The authors have declared that no competing interests exist.

advanced experimental studies of this kind in speech, in light of a systematic and computational evaluation of hypotheses concerning motor and auditory updates that could result from motor learning. To do so, we mathematically translate these hypotheses into a unified Bayesian model that integrates for the first time speech production and speech perception in a coherent architecture. We show that experimental findings are best accounted for when motor learning is assumed to generate updates of the auditory-motor internal model and the auditory characterization of phonemes, and when perception is assumed to involve both auditory and somatosensory pathways. This strongly reinforces the view that auditory and motor knowledge intervene in speech perception, and suggests likely mechanisms for motor learning in speech production.

Introduction

The fact that perception has an influence on motor learning is known and has been the focus of a large number of studies. The converse, i.e. that motor learning would influence perception, seems more intriguing and unclear. For speech, shifts in perceptual boundaries have been shown to result from motor learning induced by perturbations of the auditory feedback [1, 2] or perturbations of the articulatory gestures [3]. In the context of the well-known historical debates about the primitives (auditory/articulatory/motor) of speech perception [4–8], these findings could be interpreted as evidence in support of theories assuming the involvement of speech production processes in speech perception. However, an influence of speech motor learning on perceptual categorization of speech sounds does not necessarily imply an involvement of brain motor areas in speech perception. Indeed, the unusual auditory signals experienced during the adaptation process may by themselves be responsible for the observed perceptual shift.

From this observation, and building up on Shiller et al.'s experiment [2], Lametti et al. [1] specifically attempted to disentangle the respective influence of motor functions and altered sensory inputs on the perceptual boundary shifts. To do so, they developed an experimental protocol designed to assess separately the learning effects induced by changes in auditory feedback, on the one hand, and those arising from changes in motor control, on the other hand. They concluded that the origin of the perceptual change is indeed motor rather than sensory.

Lametti et al.'s study is very rich and relies on a solid experimental methodology. However we argue that their reasoning, because it is only qualitative, is incomplete, and does not enable to fully understand the nature of the mechanisms underlying the link observed after motor learning between changes in motor functions and perceptual changes.

In the present work we propose to dig into these questions using a previously defined Bayesian model [9]. This model was previously used to study the relative roles of auditory and proprioceptive representations in speech gesture planning; here we adapt this model to identify, implement and compare different hypotheses concerning motor adaptation. We analyze the consequences of these different hypotheses on perception and production mechanisms and suggest additional tentative interpretations of the experimental findings reported by Lametti et al. [1]. This constitutes, in our view, an important step to better relate experimental data to theories of speech production and speech perception, and further enlighten the possible role of motor processes in speech perception. Importantly, the Bayesian model we use enables to translate classical and transversal questions about motor control, perception, learning and adaptation into computations and predictions. Such a model is a methodological tool to tackle

these issues widely in speech production and speech perception, as well as in arm motor control [10, 11].

The body of this paper is divided into four sections. The remaining of this section gives an overview of the main experimental paradigms and facts reported by Lametti et al. [1]. We then present our modeling framework to deal with these experimental findings; this is presented in Section "Model". The interpretation of the results of simulations are presented in Section "Results", and discussed in Section "Discussion".

Influence of motor learning upon speech perception: Overview of experimental facts

The influence of speech motor learning on speech perception was first reported by Shiller et al. [2] (this study is called "S-09" henceforth). Motor learning was implemented by perturbing the auditory feedback of subjects when they were producing the fricative /s/: it consisted in shifting down the first spectral moment of /s/ in such a way that it sounded more like /ʃ/. They observed that subjects adapted their articulation after training in order to compensate, partially, for the perturbation, and the perceptual test after adaptation revealed a shift of the perceptual boundary between /s/ and /ʃ/ toward /ʃ/ (more sounds were perceived like /s/).

Five years later, Lametti et al. [1] published a new study (referred to as "L-14" henceforth) aiming at clarifying whether the observed perceptual change was related to "the change to motor function that occurs during learning, [to the] perceptual learning related to the altered sensory inputs, [or to] some combination of the two"(p 10339). To this end they proposed an original experimental design supposed to disentangle the effects of sensory vs. motor processes on perceptual categorization. While in S-09 a perturbation of the fricative /s/ was introduced in only one direction (toward the fricative /ʃ/), in L-14 the vowel /ɛ/ was perturbed in two directions. For one group of subjects, the perturbation was applied toward the vowel /a/ by increasing the frequency of the first formant F_1 (left panel in Fig 1). For the other group it was applied toward the vowel /i/ by decreasing F_1 (right panel in Fig 1).

To make the reasoning in L-14 clear, let us analyze the case of the perturbation toward /a/ (see Fig 1, left panel). The shift of the auditory percept along the /ɛ-a/ continuum generated a

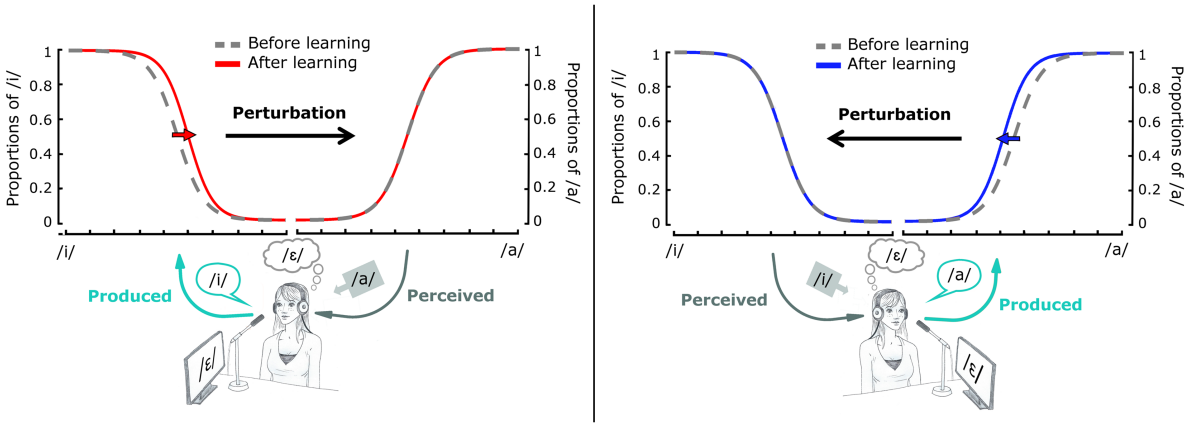

Fig 1. Illustration of results obtained by Lametti et al. [1]. Perceptual categorization curves before and after motor learning. Left panel: perturbation of vowel /ɛ/ toward vowel /a/. Right panel: perturbation of vowel /ɛ/ toward vowel /i/. Subjects compensate by producing sounds opposed to the direction of perturbation, closer to /i/ in the first case, and closer to /a/ in the second. Perceptual boundary shifts are observed for both directions of perturbations. The shift goes in the same direction as the perturbation, and is present only in the portion of auditory space corresponding to the productions of subjects during the compensation. Adapted from Lametti et al. [1].

https://doi.org/10.1371/journal.pcbi.1005942.g001

compensatory movement of the tongue frontwards, which corresponds in the absence of perturbation to an auditory percept along the /ɛ-i/ continuum. Since compensation is never complete, it results with altered auditory feedback in /ɛ/ sounds that remain partly perturbed and belong to the /ɛ-a/ region, while speaker's gestures and their corresponding somatosensory information actually belong to the /ɛ-i/ region. This is the clever method used by the authors to attempt to disentangle auditory and motor interpretations of the perturbation effects. Indeed, in their reasoning, measuring the shift of the perceptual boundary between /ɛ/ and /a/ provides a measure of the effect of the altered sensory inputs on perceptual categories, while measuring the shift of the perceptual boundary between /ɛ/ and /i/ provides a measure of the effects of the changed articulation, i.e. of the motor function, on perceptual categories. A symmetric reasoning applies for the perturbation toward /i/ (Fig 1, right panel).

Concerning motor learning, consistent with S-09 and other auditory perturbation studies in speech, motor compensation was observed and its magnitude was on average below 40% of the amplitude of perturbation. Concerning perception, a significant boundary shift was also observed in L-14. Consistent with observations reported in S-09, the resulting perceptual shifts were in the same direction as the perturbation. However, contrary to S-09, no significant shift was observed in L-14 in the region of the altered auditory inputs (i.e. the /ɛ-a/ continuum for a perturbation towards /a/); the significant shift was found in the region corresponding to the altered articulation (i.e. the /ɛ-i/ continuum for a perturbation toward /a/, see Fig 1, left panel). A control group in which subjects produced the same sequence of sounds without alteration of the auditory feedback did not show any perceptual boundary shift.

The authors concluded that their findings are "consistent with the idea that changes to central motor commands associated with speech learning are the source of changes observed in the perceptual classification of speech sounds" [1, p 10340]. Notice that if it is true that the origin of the observed perceptual shift is due to motor functions, greater changes in motor functions should induce greater changes in perception, inducing after learning positive correlations between the amount of compensation and the amplitude of the resulting perceptual shift. Intriguingly, an absence of significant correlation was reported in L-14.

Summary of experimental results we aim at modeling

Our aim is to exploit a previously defined computational framework [12–14] modelling the interactions of perception and production in speech communication, and to apply it to model and better understand the experimental data of L-14. In our modeling approach our prime concern is to extract the deeper meaning of the experimental observations and to specify a limited number of facts that best characterize them. The following summary presents the main experimental facts on which we will focus in our modeling work.

1. **Changes in speech production induced by auditory perturbations**.
    a. **Motor compensation:** speaker's articulatory movements are modified to reduce the impact of the perturbation on the perceived sound.
    b. **Incomplete compensation:** compensatory maneuvers never fully cancel the effects of the perturbation. On average, compensatory spectral changes are always below 40% of the magnitude of perturbation.
    c. **Motor adaptation:** when the perturbation is removed after the learning phase, changes in speech production remain during a certain number of trials. This so called after-effect reflects a reorganization of the motor planning process that precedes motor execution of speech gestures.

2. **Changes in speech perception**. Both motor adaptation studies, S-09 and L-14, report shifts in boundaries between phonemic perceptual categories. The key-observations are:
   a. **Consistency in the direction:** on average, across subjects, the direction of the shift is the same as the direction of the perturbation in both L-14 and S-09.
   b. **Presence of an asymmetry:** in L-14 a significant perceptual boundary shift was observed only in the portion of the auditory space related to the articulation of subjects when compensating for the perturbation, and not in the portion of auditory space related to what subjects heard in presence of the perturbation. This asymmetry was not explored in S-09 on fricative /s/ because there is no phoneme category beyond /s/ in a direction opposite to /∫/ along the spectral continuum /s-∫/. It should be noted though that the results of S-09 tend to contradict the interpretation provided in L-14 since they describe a perceptual shift in the portion of the space related to what subjects heard in the presence of the auditory perturbation.
3. **Absence of correlations between amounts of motor compensation and perceptual shift**. Both the amount of motor compensation and the amount of perceptual boundary shift differ across subjects. While one would expect a relation between the amount of compensation and the resulting perceptual shift, no significant correlation was found in L-14.

Model

This section introduces our model, which is an instance of the Bayesian algorithmic modeling framework [15], that is, the application of Bayesian Programming [16] to Marr's algorithmic level of cognitive modeling [17]. With this framework, we have previously developed a series of models, under the COSMO moniker, to study speech perception and speech production in different contexts, such as speech communication and the emergence of phonological systems [13], speech perception in adverse conditions [12, 14], sensorimotor learning [18] and the emergence of speech idiosyncrasies [19]. Variants have also been applied, in speech production, to token-to-token variability [20], the incorporation of multiple constraints in speech planning [21] and the modeling of multisensory (acoustic and somatosensory) speech targets [9]. It is this last variant that we adapt here to our current study.

In the Bayesian algorithmic modeling approach, an overarching feature is that perception and production processes are not directly modeled. Instead, we build an undirected model of speech-relevant knowledge using probability distributions. Then, from this model, we compute distributions using Bayesian inference to simulate perception and production tasks. Perception and production processes, therefore, if they involve the same knowledge, become related. Let us consider the case of speech: in our approach, we commonly assume that the description of acoustic targets in speech planning is the same piece of knowledge as would be used in a purely auditory decoder in speech perception. This distinction between the knowledge stored in the model and its use to generate processes makes our framework ideal for the study of the links between production and perception mechanisms, such as those addressed in this work.

The model includes selected aspects of speech production and speech perception that are described in Section "Selected aspects for modeling". Their implementation in the model is explained in Sections "Model definition" and "Formulation of speech production and perception questions". The strategy used to simulate the experimental paradigm of L-14 is detailed in Section "Implementation of the experimental paradigm: Normal vs. adapted conditions". Finally, the simulation results and their analysis are presented in Section "Results".

Selected aspects for modeling

Our aim is to study the interaction between speech production and speech perception processes in light of the experimental results provided in L-14. The first step in such a modeling approach consists in reducing the complexity of the experimental world into a core set of simplified components likely to capture its essential ingredients. This simplification phase should result in constraining and focusing both model implementation and results interpretation. We have selected a reduced number of aspects in speech production and speech perception that we consider to be crucial and sufficiently representative for the investigation of the interaction between motor learning and perception of isolated phonemes—here, isolated vowels /i/, /ɛ/ and /a/.

1. **Considering the stable states before and after learning**. We do not consider the particular details of the trial-to-trial evolution of the adaptation process during the training phase. Instead, we only focus on the stable states preceding and reached at the end of the adaptation process.

2. **Priority is given to speech motor planning**. We do not include any modeling of the execution of speech production gestures, ignoring in particular online feedback correction mechanisms, and only focus on the early offline planning stage preceding motor execution.

3. **Time independent states**. In the context of the two previous assumptions, we further simplify the speech production and perception systems by considering only time independent motor and sensory states that would correspond to stable vowel utterances.

4. **One-dimensional linear description**. Since both experimental designs in S-09 and L-14 studied perturbation and perception along a single dimension of the auditory space, we formally reduce the high dimensionality of motor and sensory spaces to a unique dimension. In addition, as a first order approximation, we assume that the relation between motor and sensory spaces is linear. This one-dimensional-linear simplification cannot account for the well-known many-to-one relationships between motor commands and articulatory configurations (most evident in co-contraction [22]), on the one hand, and articulatory configuration and acoustic signal on the other hand [23]. However, while this aspect would be crucial in motor learning based on articulatory perturbation (bite-block, lip-tube, jaw perturbation) requiring the use of motor-equivalence strategies for the subjects to compensate for the perturbation, it is not at the core of the mechanisms investigated in S-09 and L-14. Hence, for the sake of computational simplicity and interpretability of the results, we discard this complexity from the present analysis. This enables to take a coarse grain view and to focus on qualitative effects concerning different possible assumptions about motor adaptation, which will be introduced in Section "Implementation of the experimental paradigm: Normal vs. adapted conditions".

5. **Auditory and somatosensory properties of the sensory representations of speech units**. Finally, a fundamental question underlying the definition of our model concerns the sensory nature of speech units. Sensory representations are usually assumed to account for classification of speech sounds in perception and for the definition of motor goals in production. Concerning production, the presence of compensatory behavior induced by auditory perturbations has been a main argument supporting the hypothesis that speech motor goals are essentially characterized in auditory terms [24, 25]. However, somatosensory perturbation studies have also reported significant compensation in speech related movement, also suggesting the existence of somatosensory characterizations of speech motor goals [26, 27]. Concerning perception, auditory representations of course play a key role. This has

been confirmed for the perception of self-generated speech via perturbation experiments such as those using lip tubes or perturbation of the auditory feedback [28, 29] and it is in line with all reviews of the neuroanatomy of speech perception (e.g. [30–33]). However it remains unclear whether these are the only sensory representations that may be involved. In particular, a number of studies show an influence of somatosensory inputs on the perception of speech sounds [34] and neurocognitive data converge on the view that somatosensory regions are involved in speech processing (see a recent review by Skipper et al. [35]), suggesting a possible involvement of somatosensory representations as well. Our position with respect to these questions is the following:

a. **In production**, we assume the involvement of both auditory and somatosensory representations.

b. **In perception**, we consider two alternatives and evaluate their consequences in our framework: either perception of speech sounds involves auditory representations only or it involves both auditory and somatosensory representations. This will provide the underlying key question of this work, namely whether the data reported in L-14 do support the involvement of the speech production system in speech perception through the somatosensory system.

Model definition

The structure of the model consists in implementing a chain of probabilistic dependencies between phonological, motor and sensory variables. Variables and their dependencies are illustrated in Fig 2, and we now describe the most salient aspects of the model (a more complete mathematical description is provided in Supporting information S1 Text.

Variables. Variables in the model can be grouped into three sets. The first set is structured around variable M, which represents the set of motor commands that drive speech gestures. Associated to this variable are two "sensory-motor" variables, A_M and S_M, which represent respectively the expected auditory and somatosensory consequences of motor commands M. As stated previously, both motor and sensory-motor variables are assumed to be one-dimensional continuous variables.

The second set is structured around variable Φ, which represents the units of speech to be produced or perceived. As stated previously, we only consider vowels /i/, /ɛ/ and /a/. Associated to variable Φ are two "sensory-phonological" variables, A_Φ and S_Φ, which characterize these speech units in auditory and somatosensory terms respectively. As for sensory-motor variables, sensory-phonological variables are assumed to be one-dimensional continuous variables.

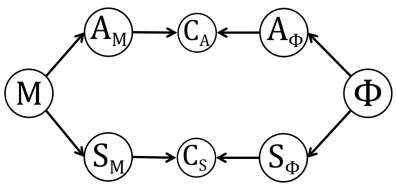

Fig 2. Graphical representation of model dependencies. Nodes represent variables and arrows display dependency relations. Variable Φ corresponds to phonemes, which are characterized in terms of auditory and somatosensory variables A_Φ and S_Φ. Variables A_M and S_M represent the predicted auditory and somatosensory consequences of the motor commands M. Variables C_A and C_S implement two sensory-matching constraints that allow the connection of the corresponding sensory pathways.

https://doi.org/10.1371/journal.pcbi.1005942.g002

The last set of variables link sensory-phonological and sensory-motor variables. C_A and C_S are two coherence variables, which are "probabilistic connectors" between variables A_M and A_Φ, for C_A, and between variables S_M and S_Φ, for C_S. These connectors can be either left "open", in which case the variables they link are mathematically independent, or "closed", in which case the variables they link are forced to have the same value by a matching constraint. As such, these coherence variables can be interpreted as a "mathematical trick" to implement Bayesian switches [16, 36] controlling the propagation of information in the model.

Dependencies: Decomposition of the joint probability distribution. The joint probability distribution is decomposed as a product of elementary terms:

$$\begin{aligned} & P(M\ S_M\ A_M\ \Phi\ S_\Phi\ A_\Phi\ C_S\ C_A) \\ & = P(M)P(A_M \mid M)P(S_M \mid M) \\ & \quad P(\Phi)P(A_\Phi \mid \Phi)P(S_\Phi \mid \Phi) \\ & \quad P(C_A \mid A_M\ A_\Phi)P(C_S \mid S_M\ S_\Phi). \end{aligned} \tag{1}$$

This decomposition, illustrated in Fig 2, relies on a certain number of conditional independence hypotheses that we do not discuss here (but see Supporting information S1 Text for details).

Parametric forms. We now define each probability distribution of Eq (1). Concerning prior distributions $P(M)$ and $P(\Phi)$, we assume no prior knowledge concerning values of variables M and Φ. Therefore, we identify $P(M)$ and $P(\Phi)$ with uniform distributions.

$P(A_M \mid M)$ and $P(S_M \mid M)$ represent knowledge relating motor commands to their predicted sensory consequences. They correspond to sensory-motor internal forward models often assumed to be involved in motor planning [37–39] (but see [40, 41] for debates). As explained in Section "Selected aspects for modeling", for the sake of computational simplicity, we assume that these stored relations are linear. The corresponding auditory-motor and somatosensory-motor mappings, $\rho_A(m)$ and $\rho_S(m)$, are defined as follows:

$$\rho_A(m) \quad \coloneqq \quad \alpha_A.m + \beta_A, \tag{2}$$

$$\rho_S(m) \quad \coloneqq \quad \alpha_S.m + \beta_S, \tag{3}$$

where values of parameters α_A, α_S, β_A and β_S depend on further hypotheses that will be specified in Section "Implementation of the experimental paradigm: Normal vs. adapted conditions". Finally, we further assume that the stored sensory-motor internal models have infinite precision and are therefore deterministic, such that $P(A_M \mid M)$ and $P(S_M \mid M)$ are identified with Dirac delta functions:

$$P([A_M = a] \mid [M = m]) \quad \coloneqq \quad \delta(a - \rho_A(m)), \tag{4}$$

$$P([S_M = s] \mid [M = m]) \quad \coloneqq \quad \delta(s - \rho_S(m)). \tag{5}$$

$P(A_\Phi \mid \Phi)$ and $P(S_\Phi \mid \Phi)$ correspond to the auditory and somatosensory characterizations of phonemes. As it is common in other modeling studies [42–46], we identify them with Gaussian distributions specified by their means and standard-deviations $(\mu_A^\phi, \sigma_A^\phi)$ and $(\mu_S^\phi, \sigma_S^\phi)$ for each phoneme ϕ in auditory and somatosensory terms. Values of parameters $(\mu_A^\phi, \sigma_A^\phi)$ and $(\mu_S^\phi, \sigma_S^\phi)$ depend on further hypotheses and will be specified in Section "Implementation of the experimental paradigm: Normal vs. adapted conditions" and "Update of the auditory-motor internal model $P(A_M \mid M)$".

$P(C_A \mid A_M\, A_\Phi)$ and $P(C_S \mid S_M\, S_\Phi)$ implement the sensory matching constraints relating sensory-motor and sensory-phonological variables in the following way:

$$P([C_A = 1] \mid [A_M = a_m]\ [A_\Phi = a_\phi]) \quad \coloneqq \quad \begin{cases} 1 & \text{if } a_m = a_\phi \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$P([C_S = 1] \mid [S_M = s_m]\ [S_\Phi = s_\phi]) \quad \coloneqq \quad \begin{cases} 1 & \text{if } s_m = s_\phi \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Formulation of speech production and perception questions

In the previous section we proposed a computational definition of the joint probability distribution of the model. This definition was based on particular assumptions concerning relations between variables. The Bayesian formalism allows to simulate speech production and perception by defining and computing probability distributions of interest, that we call "questions".

Speech production questions. Speech production questions correspond to the inference of motor commands for the production of a desired phoneme. The dependence structure of Fig 2 shows that if coherence variables C_A and C_S are not assumed to be 1, that is to say, if they are "Bayesian switches" left open, there is no dependency between M and Φ, which would correspond to an unrealistic situation (see Supporting information S2 Text for further details). Instead, inferring motor commands for the production of a given phoneme with either variable C_A or variable C_S or both set to 1, leads to three planning processes that can be characterized as follows.

1. The first planning process is based on the auditory pathway only and corresponds to:

$$\begin{aligned} &P([M = m] \mid \Phi\ [C_A = 1]) \\ &\propto P([A_\Phi = \rho_A(m)] \mid \Phi). \end{aligned} \tag{8}$$

2. The second planning process is based on the somatosensory pathway only and corresponds to:

$$\begin{aligned} &P([M = m] \mid \Phi\ [C_S = 1]) \\ &\propto P([S_\Phi = \rho_S(m)] \mid \Phi). \end{aligned} \tag{9}$$

3. The third planning process is based on the fusion of auditory and somatosensory pathways and corresponds to:

$$\begin{aligned} &P([M = m] \mid \Phi\ [C_A = 1]\ [C_S = 1]) \\ &\propto P([A_\Phi = \rho_A(m)] \mid \Phi) P([S_\Phi = \rho_S(m)] \mid \Phi), \end{aligned} \tag{10}$$

These equations are obtained by the application of Bayesian inference rules to the joint probability distribution given by Eq (1). Derivations are provided in Supporting information S2 Text. All terms on the right hand sides of Eqs (8), (9) and (10) were defined in Section "Parametric forms".

The probability of selecting a particular motor command m is hence proportional to the probability that the predicted sensory consequences of m (expressed by $\rho_A(m)$ and $\rho_S(m)$ in auditory and somatsoensory terms) are in agreement with the sensory characterization of the intended phoneme in the corresponding sensory pathway.

Speech perception questions. Perception questions correspond to the categorization of auditory inputs into phoneme identity. We consider that the perceived auditory stimulus is a value of the auditory-motor variable A_M. Similar to the previous production questions, we can define three categorization questions depending on the activation of variables C_A or C_S.

1. The assumption that categorization is based only on the auditory pathway (as in auditory theories of speech perception) corresponds to:

$$P([\Phi = \phi] \mid [A_M = a]\ [C_A = 1]) = \frac{P([A_\Phi = a] \mid [\Phi = \phi])}{\sum_{\phi'} P([A_\Phi = a] \mid [\Phi = \phi'])} \tag{11}$$

2. The assumption that categorization is based only on the somatosensory pathway (as in the direct realist theory [47]) corresponds to:

$$P([\Phi = \phi] \mid [A_M = a]\ [C_S = 1]) = \frac{P([S_\Phi = \rho_S \circ \rho_A^{-1}(a)] \mid [\Phi = \phi])}{\sum_{\phi'} P([S_\Phi = \rho_S \circ \rho_A^{-1}(a)] \mid [\Phi = \phi'])} \tag{12}$$

3. The assumption that categorization is based on the fusion of both auditory and somatosensory pathways (as in perceptuo-motor theories [8]) corresponds to:

$$P([\Phi = \phi] \mid [A_M = a]\ [C_S = 1]\ [C_A = 1]) = \frac{P([A_\Phi = a] \mid [\Phi = \phi])\ P([S_\Phi = \rho_S \circ \rho_A^{-1}(a)] \mid [\Phi = \phi])}{\sum_{\phi'} P([A_\Phi = a] \mid [\Phi = \phi'])\ P([S_\Phi = \rho_S \circ \rho_A^{-1}(a)] \mid [\Phi = \phi'])}. \tag{13}$$

The symbol $\circ$ in Eqs (12) and (13) denotes the composition operator, and therefore $\rho_S \circ \rho_A^{-1}(a)$ corresponds to the somatosensory image of the auditory value a as obtained first by the identification of motor commands m achieving the production of a ($m = \rho_A^{-1}(a)$) and then by the prediction of the somatosensory variable s generated from the inferred motor commands ($s = \rho_S(m)$). Solutions for these three inference questions are obtained from the joint probability distribution given by Eq (1). Details of the derivation are provided in Supporting information S2 Text.

These equations express the way Bayesian computation yields categorization processes from the structure and knowledge encoded in the model. Under the auditory pathway case, the probability of categorizing an auditory input a into phoneme ϕ is obtained by evaluating the probability that this auditory input would correspond to the auditory characterization of the considered phoneme ($P([A_\Phi = a] \mid [\Phi = \phi])$ in the numerator), and comparing it to the probability that it would correspond to the auditory characterization of any of the possible phonemes (the sum over ϕ' on the denominator). When this ratio is close to 1, the auditory value is categorized as phoneme ϕ with full certainty. The smaller the ratio, the lower the probability of this categorization.

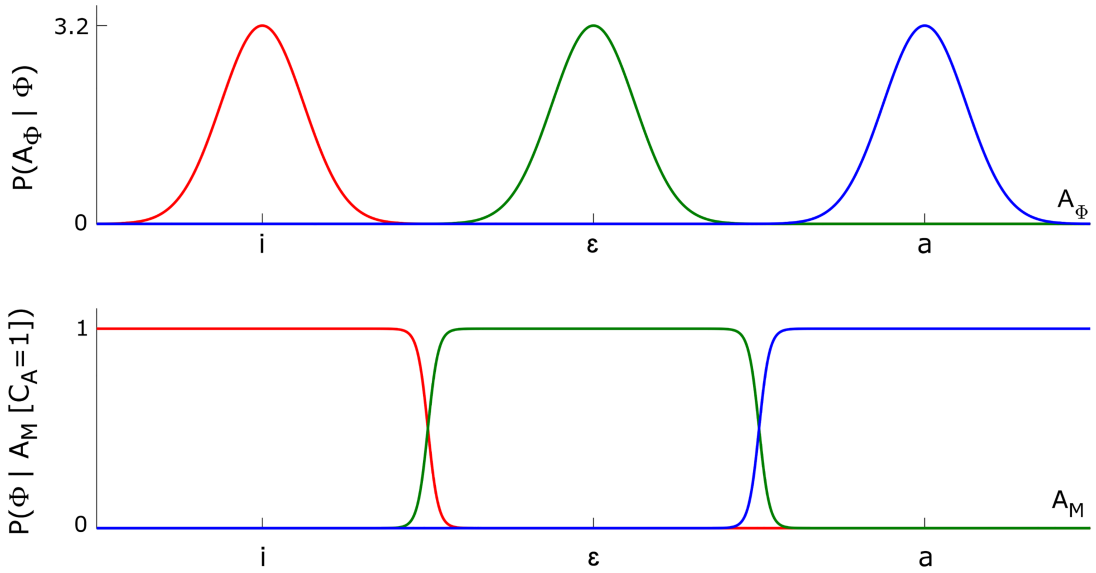

Fig 3. Auditory characterizations and corresponding phoneme categorization functions. Top panel: auditory characterization, $P(A_\Phi \mid \Phi)$, for phoneme /i/ (red), phoneme /ε/ (green) and phoneme /a/ (blue). Bottom panel: categorization functions under the auditory pathway approach, $P(\Phi \mid A_M\,[C_A = 1])$, obtained from auditory characterizations according to Eq (11). Eq (14) gives the explicit form for phoneme /i/.

https://doi.org/10.1371/journal.pcbi.1005942.g003

Consider, for instance, the case of the categorization of an auditory input *a* into the phoneme /i/ (among the three vowels /i, ε, a/ in our example) as given by Eq (11). Replacing $P(A_\Phi \mid \Phi)$ with their definition as Gaussian probability distributions yields:

$$P([\Phi = /\mathrm{i}/] \mid [A_M = a]\,[C_A = 1]) = \frac{e^{-\frac{(a-\mu_A^i)^2}{2\sigma_A^{i\,2}}}}{e^{-\frac{(a-\mu_A^i)^2}{2\sigma_A^{i\,2}}} + e^{-\frac{(a-\mu_A^\varepsilon)^2}{2\sigma_A^{\varepsilon\,2}}} + e^{-\frac{(a-\mu_A^a)^2}{2\sigma_A^{a\,2}}}}. \quad (14)$$

This function is illustrated in Fig 3 for parameter values in the normal condition, as specified in Section "Normal condition: Initial values of parameters". The corresponding categorization functions under the somatosensory and fusion of pathways are derived essentially in the same way.

Selection of production and perception questions. We have derived 3 perception and 3 production questions that differ with respect to the sensory pathways assumed to be involved in these processes. For the sake of brevity, we limit the presentation of simulations and do not consider the outcome of all of the 9 combinations of questions, in order to focus on those that correspond to the richest scientific contributions. Therefore, as pointed out in Section "Selected aspects for modeling", concerning production we consider only the question assuming the fusion of auditory and somatosensory pathways, $P(M \mid \Phi\,[C_A = 1]\,[C_S = 1])$. Concerning perception we keep and compare questions assuming the involvement of the auditory pathway alone, $P(\Phi \mid A_M\,[C_A = 1])$, and the fusion of sensory pathways, $P(\Phi \mid A_M\,[C_A = 1]\,[C_S = 1])$.

In order to simplify notations, we denote the selected production and perception questions by:

$$\begin{aligned} Q^{F}_{\text{Prod}} &\coloneqq P(M \mid \Phi\ [C_A = 1]\ [C_S = 1]), \\ Q^{A}_{\text{Per}} &\coloneqq P(\Phi \mid A_M\ [C_A = 1]), \\ Q^{F}_{\text{Per}} &\coloneqq P(\Phi \mid A_M\ [C_A = 1]\ [C_S = 1]). \end{aligned}$$

Implementation of the experimental paradigm: Normal vs. adapted conditions

Our aim is to simulate and compare the outcome of the production and perception tests in L-14, prior to the auditory perturbation and after the training phase, i.e. when perturbation is removed and adaptation has been reached. These tests are naturally implemented in the model as the outcome of the production and perception questions defined in the previous section.

Adaptation is implemented as the update of a part of the knowledge included in the model. This knowledge is represented by the four relations defined in Section "Parametric forms": the two sensory-motor internal models, $P(A_M \mid M)$ and $P(S_M \mid M)$, and the two sensory characterizations of phonemes, $P(A_\Phi \mid \Phi)$ and $P(S_\Phi \mid \Phi)$. In this context, normal and adapted conditions are implemented by different values of the parameters characterizing these relations. Values of parameters in normal condition are arbitrary initial values. This is why we chose them to be as simple as possible. They are specified in Section "Normal condition: Initial values of parameters".

Two fundamental questions remain to be answered in order to specify how adaptation will affect these initial values: (1) which of the four relations is changed during adaptation, and (2) how? The first question actually rephrases in computational terms the question raised in L-14 (p 10339), and quoted in its original formulation in Section "Influence of motor learning upon speech perception: overview of experimental facts", extending it to behavioral changes in both production and perception: "So what produces the [behavioral changes] during motor learning? Is it the change to [parameters of the sensory-motor internal models], that occurs during learning? Is it changes to [parameters of the sensory characterizations of phonemes], related to the altered sensory inputs? Or is it some combination of the two?".

In the following sections, we address these two questions in two steps. In Section "Adaptation hypotheses" we partially answer the first question by motivating the selection of a subset of possible changes induced by adaptation. In Section "Results" we further answer these questions by evaluating the outcome of different implementations of the selected changes and by comparing them with the experimental facts summarized in Section "Summary of experimental results we aim at modeling".

Normal condition: Initial values of parameters. We now specify parameters of the two sensory-motor internal models, $P(A_M \mid M)$ and $P(S_M \mid M)$, and the two sensory characterizations of phonemes, $P(A_\Phi \mid \Phi)$ and $P(S_\Phi \mid \Phi)$.

The two sensory-motor internal models are defined in terms of auditory-motor and somatosensory-motor mappings ρ_A and ρ_S, which are characterized by parameters α_A, β_A and α_S, β_S. Without loss of generality, we define metric units of motor and sensory spaces in order to have $\alpha_A = \alpha_S = 1$ and $\beta_A = \beta_S = 0$ in normal condition. Therefore, with ρ^n_A and ρ^n_S being the auditory-motor and somatosensory-motor mappings in normal conditions respectively, we have:

$$\rho^{(n)}_A(m) = m, \tag{15}$$

$$\rho^{(n)}_S(m) = m. \tag{16}$$

The left panels of Fig 4 illustrate these sensory-motor mappings.

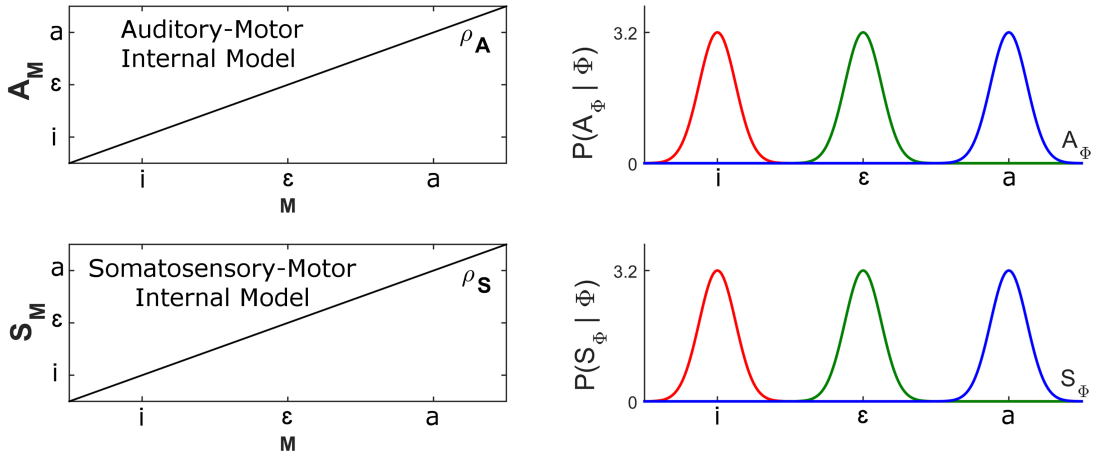

Fig 4. Stored sensory-motor mappings and sensory characterizations of phonemes under normal conditions. Left panels: auditory-motor internal mapping ρ_A^n (top) and somatosensory-motor internal mapping ρ_S^n (bottom). Both mappings are assumed to be identity. Right panels: auditory (top) and somatosensory (bottom) characterization of phonemes. Probability distributions are all Gaussian, evenly distributed in each space, and with equal standard-deviations, equal to $\frac{1}{8}$ of the distance between phonemes.

https://doi.org/10.1371/journal.pcbi.1005942.g004

The two sensory characterizations of phonemes are defined as Gaussian probability distributions with means and standard-deviations, $(\mu_A^\phi, \sigma_A^\phi)$ and $(\mu_S^\phi, \sigma_S^\phi)$.

For the sake of simplicity, we assume that in normal condition these sensory characterizations are evenly distributed in both sensory spaces with the same standard-deviations equal to $\frac{1}{8}$ of the distance between neighboring phonemes (see Supporting information S4 Text for further details). The right panels of Fig 4 illustrate the corresponding probability distributions.

Since the model is now completely defined in normal conditions, we can study the outcome of the production and perception questions, which correspond to production and perception pretests in L-14. The corresponding functions are displayed in Fig 5.

Concerning production, Eq (10) indicates that the outcome of the planning process is a product of two Gaussian probability distributions. The product is known to result into a new Gaussian probability distribution with smaller variance [48], as it can be seen in the left panel displayed in Fig 5.

Concerning perception, the outcome of the two perception processes corresponds to categorization functions with the same positions of the boundaries, but with boundary slopes that are different. In this context, the fusion of sensory pathways results in a steeper slope than the auditory pathway alone.

Adaptation hypotheses. We focus now on the adapted state. Which of the two sensory-motor internal models, $P(A_M \mid M)$ and $P(S_M \mid M)$, or the two sensory characterizations of phonemes, $P(A_\Phi \mid \Phi)$ and $P(S_\Phi \mid \Phi)$ is being updated during the training phase? We consider that any of these relations may be updated if the perturbation introduced during the training phase leads to considering that they are no longer correct.

Since the perturbation of the auditory feedback only affects the relation between motor commands and auditory outputs, we do not introduce any change to the somatosensory-motor internal model, $P(S_M \mid M)$, but we assume that the auditory-motor internal model, $P(A_M \mid M)$, may be updated in order to learn the new auditory-motor relation.

The auditory perturbation also induces a mismatch between the perturbed auditory output and the learned phoneme characterization $P(A_\Phi \mid \Phi)$. This mismatch can be resolved by an update of the auditory-motor internal model alone, $P(A_M \mid M)$, so that under the perturbed

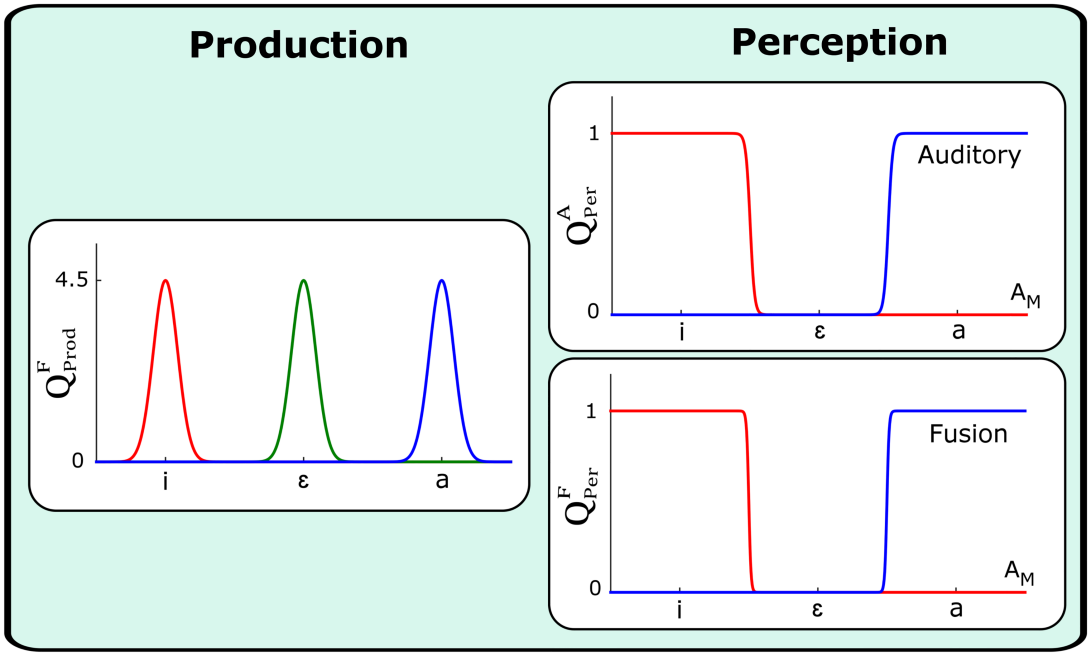

Fig 5. Outcome of the production question (left panel) and perception questions (right panels) under normal conditions. The categorization function corresponding to vowel /ɛ/ is not represented for clarity of the figure and since it corresponds to the complementary of the two other curves (as it can be seen in the bottom panel of Fig 3).

https://doi.org/10.1371/journal.pcbi.1005942.g005

condition new motor commands are associated to the usual auditory region characterizing the produced phoneme. However, modifying the auditory characterization of phonemes, $P(A_\Phi \mid \Phi)$, may also contribute to the reduction of this mismatch. In S-09 the reported results were interpreted as a combination of these two hypotheses. The authors suggested that "speech adaptation to altered auditory feedback is not limited to the motor domain, but rather involves changes in both motor output and auditory representations of speech sounds that together act to reduce the impact of the perturbation" [2, p 1103, abstract]. In other words, subjects could reduce the impact of the perturbation by changing the motor commands associated to the production of the phoneme, but also by modifying their stored auditory characterizations of speech sounds. Furthermore, as in L-14 we only focus on the perturbation of vowel /ɛ/. Hence, among the stored auditory characterizations we consider that only $P(A_\Phi \mid [\Phi = /\varepsilon/])$, corresponding to vowel /ɛ/, may be updated.

The motor command change resulting from the compensation for the auditory perturbation also induces a somatosensory mismatch. Indeed, somatosensory values resulting from compensation deviate from the stored somatosensory characterization of the intended phoneme. Therefore, once the compensation for the auditory perturbation starts to be efficient, the stored somatosensory characterization of the intended phoneme, $P(S_\Phi \mid \Phi)$, may also change in order to match the new somatosensory patterns associated with the modified articulation. Furthermore, since we only focus on the perturbation of vowel /ɛ/, we consider that among the stored somatosensory characterizations only $P(S_\Phi \mid [\Phi = /\varepsilon/])$ may be updated.

In summary, we retain three possible changes that may be induced by motor adaptation:

1. An update of the auditory-motor internal model $P(A_M \mid M)$;
2. An update of the auditory characterization of the perturbed vowel $P(A_\Phi \mid [\Phi = /\varepsilon/])$;

3. An update of the somatosensory characterization of the perturbed vowel $P(S_\Phi \mid [\Phi = /\varepsilon/])$.

These three possible changes result in 7 possible adaptation hypotheses, depending on whether we combine one, two or the three of them.

Section "Results" aims to evaluate which of these adaptation hypotheses may account for the experimental facts reported in L-14. This evaluation is performed by comparing the consequences of each hypothesis with respect to compensation and perceptual boundary shift as reported in Section "Summary of experimental results we aim at modeling".

We assess the direction and amount of compensation via the displacement of the motor planning distribution Q^F_{Prod} associated with /ε/ in the motor command space. We evaluate the amount of perceptual boundary shift via the displacement of the point where the categorization function Q^A_{Per} or Q^F_{Per} takes value $\frac{1}{2}$.

Finally, since the behavior of the model is symmetric around vowel /ε/, we focus only on the case of a perturbation in the direction of vowel /a/ (left panel of Fig 1). All simulations are therefore performed assuming a perturbation with a magnitude of 40% of the distance between neighboring phonemes and in the direction of vowel /a/.

Results

The primary goal of this section is to evaluate which of the 7 adaptation hypotheses account for the experimental facts reported in L-14. To do so, we proceed sequentially: we first focus on perception and evaluate results corresponding to the two categorization questions Q^A_{Per} and Q^F_{Per}. For the hypotheses that are compatible with the perceptual boundary shift observed in L-14, the associated compensation in production is evaluated, and again only the hypotheses that are compatible with the results of L-14 are kept. Finally, in a third step, we further evaluate the selected adaptation hypotheses with respect to the corresponding correlations between the amount of compensation in production and the magnitude of perceptual boundary shift.

Evaluation with respect to perception

Update of the auditory-motor internal model $P(A_M \mid M)$. We begin by considering the consequences of an update of the auditory-motor internal forward model $P(A_M \mid M)$ (see Fig 6) characterized by the mapping $\rho_A(m)$ and parameters α_A and β_A (see Eq (2)).

Since the perturbation corresponds to a constant shift δ_A in auditory space, a straightforward update of the auditory-motor mapping induced by training under the perturbed condition, $\rho^{(u)}_A$, can be obtained from the mapping in normal condition, $\rho^{(n)}_A$, as:

$$\rho^{(u)}_A(m) = \rho^{(n)}_A(m) + \delta_A = m + \delta_A, \tag{17}$$

that is to say, parameter α_A in Eq (2) remains unchanged (value 1) and β_A is updated by the amount of shift δ_A.

This first implementation corresponds to a general update of the internal model, which is not limited to the domain of variation of the motor commands experienced by the subject during the perturbation experiment. Assuming such a generalization is a strong hypothesis that has been questioned in different experimental studies (including in speech [49, 50], and in arm movements [51]). Consequently, we also consider a second, more local, update of the internal model that is limited to the range of motor commands experienced by the subject when speaking with the perturbation. We will compare the predictions of these two assumptions on the two perception questions.

We begin by considering the general update hypothesis. The left panels of Fig 6 present the outcome for the perception questions, under the auditory pathway hypothesis Q^A_{Per} and under

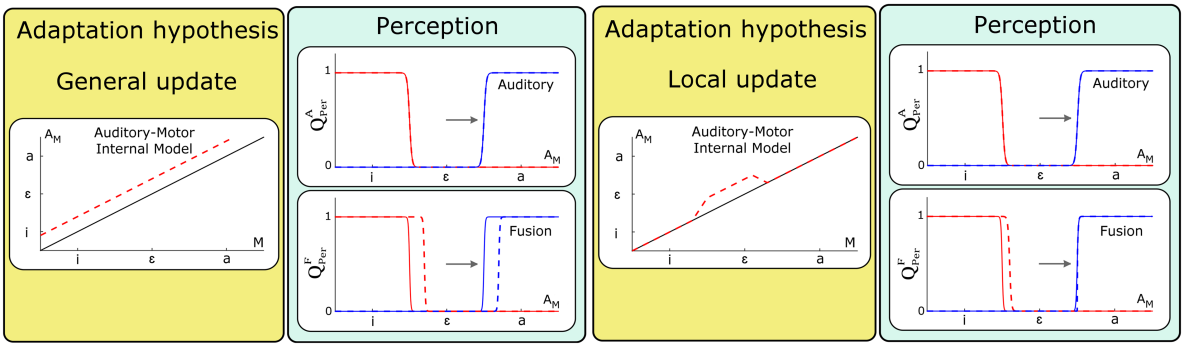

Fig 6. Changes in perception questions resulting from updates of the auditory-motor internal model $P(A_M \mid M)$. Left panels: general update. Right panels: local update. Yellow panels represent the auditory-motor mapping before (solid lines) and after update (dashed lines). The magnitude and direction of update is equal to the assumed perturbation magnitude. Green panels represent the outcome of the two perception questions in normal (solid lines) vs. adapted conditions (dashed lines). Only categorization functions for vowel /i/ (red plots) and /a/ (blue plots) are displayed (categorization function for vowel /ɛ/ being complementary to the two others). The direction and magnitude of the perturbation is indicated by the horizontal arrow.

https://doi.org/10.1371/journal.pcbi.1005942.g006

the fusion of sensory pathways hypothesis Q^F_{Per}, assuming the general update of the auditory-motor internal model $P(A_M \mid M)$. We firstly observe that updating the auditory-motor internal model results in no change in the categorization process under the auditory pathway Q^A_{Per}, consistent with Eq (11) in which only $P(A_\Phi \mid \Phi)$ is involved.

In addition, we observe a perceptual boundary shift under the fusion of pathways Q^F_{Per}. This is consistent with Eq (13) where the auditory categorization under the fusion of sensory pathways involves the inverse of the auditory-motor mapping, ρ_A^{-1}, that has been updated. Importantly, it should be noted that the direction of perceptual boundary shifts (from the solid to the dotted line) is the same as the direction of the perturbation (horizontal arrow). This is consistent with the findings reported in S-09 and L-14. However, we notice that boundary shifts are present on both sides of vowel /ɛ/ in the auditory space, contrary to the asymmetry reported in L-14.

Let us consider now the local update hypothesis. The right panels of Fig 6 present the outcome for the perception questions, Q^A_{Per} and Q^F_{Per}, assuming a local update of the auditory-motor internal model $P(A_M \mid M)$. Details about the specification of this local update are provided in Supporting information S3 Text.

The main results are consistent with those of the general update, except that the perceptual boundary shift observed under the fusion of pathways is now restricted to the region of the auditory space associated with the interval of the motor commands space where the internal model was updated, i.e. in the domain located between /i/ and /ɛ/. The resulting asymmetry is in agreement with the observations reported by L-14. However, it is important to specify that the magnitude of the shift as well as the characteristics of the asymmetry are sensitive to the choice of the parameters determining the local update of the internal model. Here, parameters implement the hypothesis that learning is limited to a portion of motor space consistent with what subjects may have explored when speaking with the perturbation.

Update of the auditory characterization $P(A_\Phi \mid \Phi)$. The auditory characterization of phonemes, $P(A_\Phi \mid \Phi)$, was identified, in Section "Parametric forms", with Gaussian distributions with parameters $(\mu_A^\phi, \sigma_A^\phi)$, where ϕ indicates the considered phoneme. In the present case, we are interested in the auditory characterization of phoneme /ɛ/, that is, in the Gaussian distribution $P(A_\Phi \mid [\Phi = /\varepsilon/])$ with parameters $(\mu_A^\varepsilon, \sigma_A^\varepsilon)$. We consider adaptation to the auditory perturbation that moves /ɛ/ toward /a/ and assume that it may update either μ_A^ε or σ_A^ε or

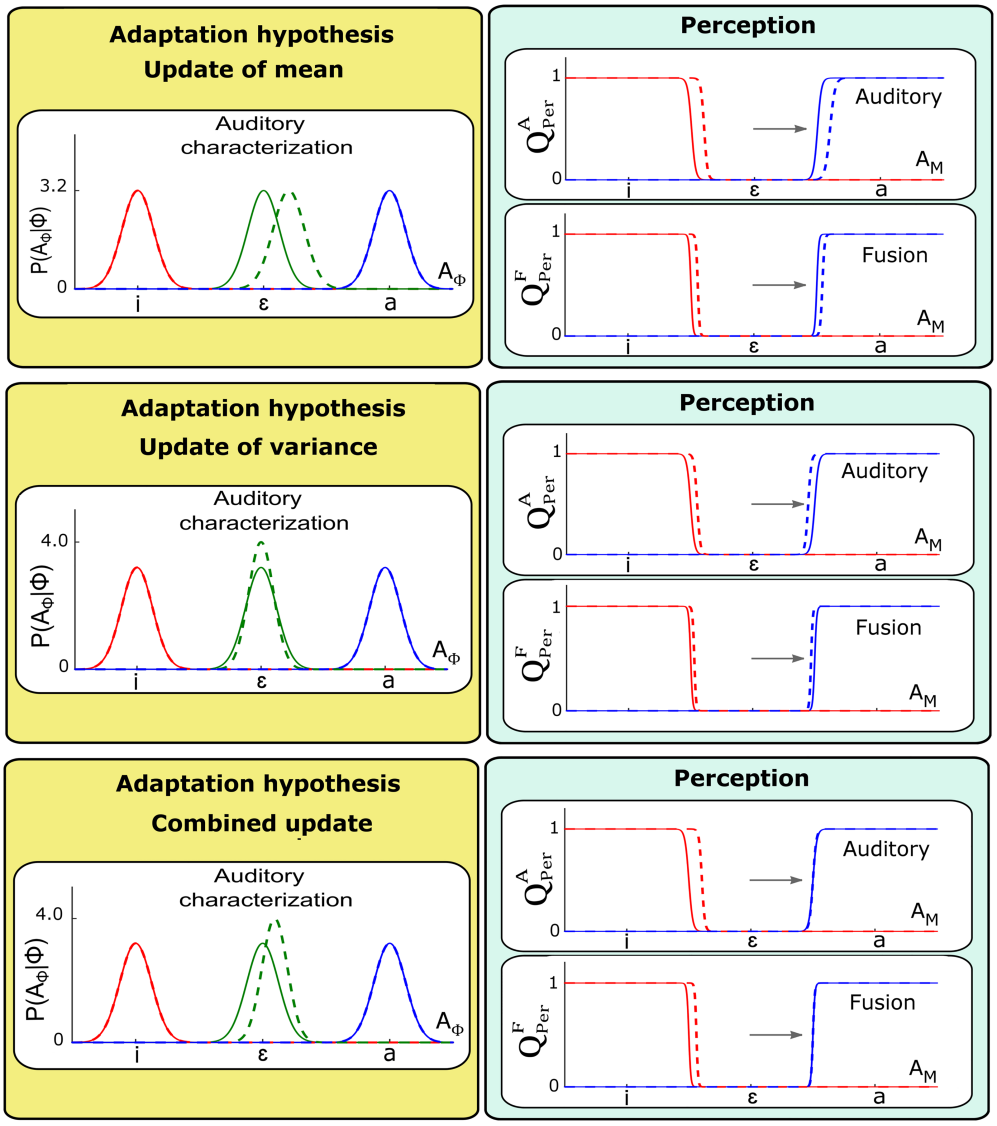

Fig 7. Changes in perception questions resulting from different updates of the auditory characterization of the perturbed vowel, $P(A_\Phi \mid [\Phi = /\varepsilon/])$. Top and middle panels correspond to independent updates of the mean μ_A^ε and the standard deviation σ_A^ε respectively. Bottom panels correspond to a combined shift of μ_A^ε and reduction of σ_A^ε. The outcome of simulations before update (solid lines) are superimposed to those after update (dashed lines).

https://doi.org/10.1371/journal.pcbi.1005942.g007

both. Updating μ_A^ε modifies the location of the Gaussian, whereas updating σ_A^ε modifies its width. We will first evaluate the effect induced by an update of each parameter independently, and then consider a combined update.

The top panel of Fig 7 present the outcome of the two perception questions Q^A_{Per} and Q^F_{Per} after a shift of μ_A^ε in the direction of vowel /a/. The middle panels of Fig 7 illustrate the outcome of the two perception questions resulting from a reduction of σ_A^ε.

We observe that modifying parameters μ_A^ε and σ_A^ε induces changes in auditory categorization both with the auditory pathway only Q^A_{Per} and with the fusion of sensory pathways Q^F_{Per}. However, the perceptual changes vary according to the parameter that is modified. Shifting parameter μ_A^ε (top panel) induces a shift of $P(A_\Phi \mid [\Phi = /\varepsilon/])$, resulting in a boundary shift that

is similar on both sides of /ε/ along the auditory continuum and goes in the same direction as the shift in location of $P(A_\Phi \mid [\Phi = /\varepsilon/])$. Reducing parameter σ_A^ε of $P(A_\Phi \mid [\Phi = /\varepsilon/])$ (middle panel) induces boundary shifts that are in opposite direction on both sides of /ε/ along the auditory continuum. The boundaries follow the narrowing of $P(A_\Phi \mid [\Phi = /\varepsilon/])$ on both sides, and get closer to the center of the Gaussian distribution characterizing /ε/.

Therefore, it appears that an adequate combination of μ_A^ε and σ_A^ε modifications may produce a pattern in agreement with the one observed by L-14, with a boundary shift in the direction of the perturbation, obtained just on the /ε/-/i/ side but not on the /ε/-/a/ side (see Fig 7, bottom panel). The relation between μ_A^ε and σ_A^ε implemented in the simulations of Fig 7 is provided in Supporting information S4 Text. Note that this relation has been specifically designed in order to reproduce the desired asymmetrical boundary shift, but we attach no claim of cognitive plausibility to this specific relation. We will discuss the theoretical implication of this ad-hoc choice in Section "Discussion".

Update of the somatosensory characterization $P(S_\Phi \mid \Phi)$. The somatosensory characterization of phonemes $P(S_\Phi \mid \Phi)$ was identified with Gaussian distributions parameterized by $(\mu_S^\phi, \sigma_S^\phi)$. The articulatory changes enabling compensation for the perturbation during adaptation could generate an update of the somatosensory characterization of the produced phoneme in order to account for the somatosensory correlates corresponding to the new articulatory postures. For an auditory perturbation of vowel /ε/ toward /a/, the compensatory behavior leads to articulatory configurations closer to /i/. One would therefore expect a change of μ_S^ε in the direction of phoneme /i/.

Fig 8 presents the outcome of the two perception questions Q_{Per}^A and Q_{Per}^F after a shift of μ_S^ε in the direction of /i/. Plots are organized in the same manner as in previous Figures. The left panel illustrates the shift in location of the somatosensory characterization of phoneme /ε/ after training, once adaptation is achieved. We observe that the update of $P(S_\Phi \mid \Phi)$ does not induce any change in perceptual categories under the auditory pathway hypothesis Q_{Per}^A. This is consistent with Eq (11) where the somatosensory characterization $P(S_\Phi \mid \Phi)$ is not involved. However, we observe a shift in auditory categorization under the fusion of pathways hypothesis Q_{Per}^F. This again is consistent with Eq (13) where the somatosensory characterization term $P(S_\Phi \mid \Phi)$ is involved. It must be noted though, that the direction of the perceptual shift is opposite to the perturbation, contrary to the experimental findings reported in S-09 and L-14.

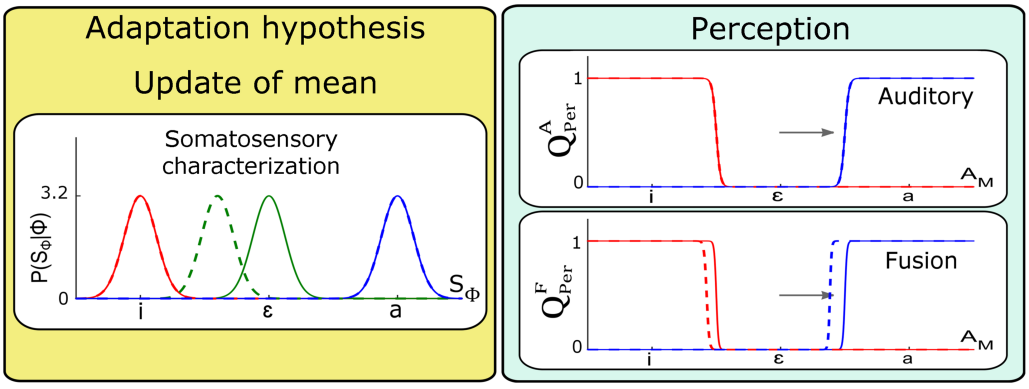

Fig 8. Changes in perception questions resulting from update of the somatosensory characterization of the perturbed vowel $P(S_\Phi \mid [\Phi = /\varepsilon/])$. The update corresponds to a shift of the mean value μ_S^ε in the direction of phoneme /i/, as would result from compensation to an auditory perturbation towards /a/ (horizontal arrow). The outcome of simulations before update (solid lines) are superimposed to those after update (dashed lines).

https://doi.org/10.1371/journal.pcbi.1005942.g008

Effects of combined update hypotheses. Until now we have only considered individual update hypotheses. The conclusion of these individual evaluations is that the local update of the auditory-motor internal model and the coordinated shift and narrowing of the auditory characterization of phoneme /ɛ/ are the only updates that correctly account for the experimental observations (direction of the shift of the boundaries between perceptual categories and asymmetry of the magnitude of the shift on both sides of vowel /ɛ/) in L-14. These updates are not exclusive and they could be involved simultaneously during adaptation. Hence, in this section we investigate the consequence for the perception questions of the combination of these two updates. Note that we are not discarding an additional update of the somatosensory characterization of phonemes in combination with the two other ones. It could be actually involved and result in a reduction of the perceptual shift induced by any of the two other hypotheses. Since this would only act as an amplification/reduction factor of the main phenomenon, we do not consider it in the remaining of this study.

Fig 9 presents the outcome of the two perception questions Q^{A}_{Per} and Q^{F}_{Per} after combination of these update hypotheses. Plots are organized in the same manner as in previous Figures. The two left panels illustrate the changes of the auditory characterization of phoneme /ɛ/ (top) and the auditory-motor mapping (bottom). Consistent with the results presented in Figs 6 and 7, we observe that after these two combined updates both the auditory and the sensory fusion accounts of perception, Q^{A}_{Per} and Q^{F}_{Per}, result in asymmetric perceptual shifts. These shifts go in the direction of the auditory perturbation and are visible only in the portion of auditory space located with respect to vowel /ɛ/ on the opposite side of the region in which the auditory perturbation was applied, in agreement with the experimental findings reported in L-14.

Summary. In summary, based on the results of the simulations presented in this section, we select 3 adaptation hypotheses that, combined with at least one of the two perception

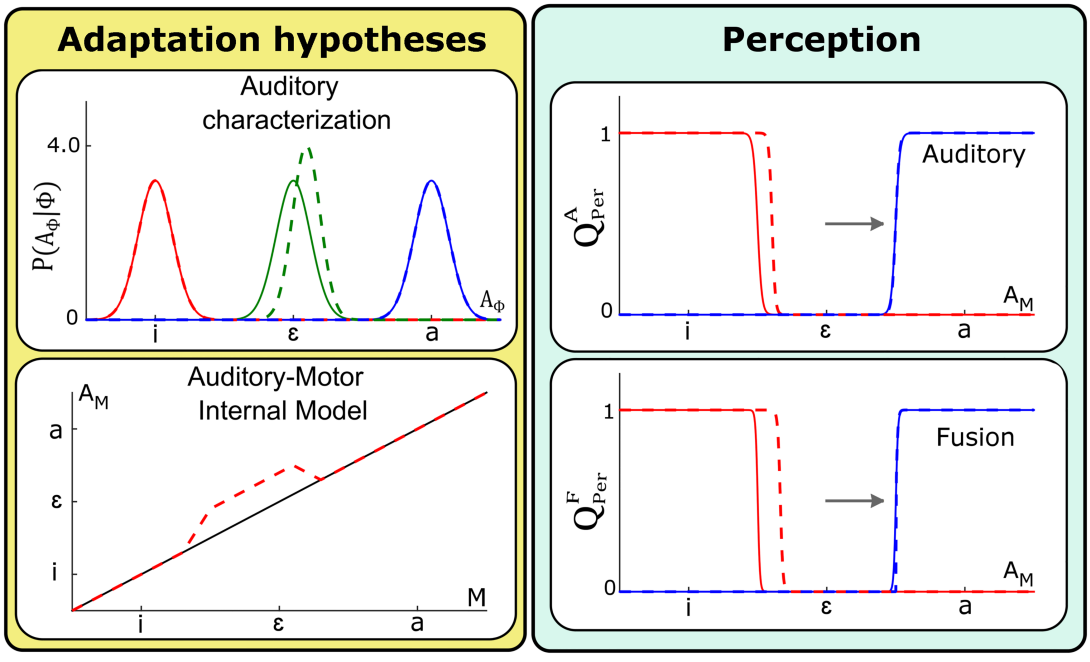

Fig 9. Changes in perception questions resulting from a combined local update of the internal model and an update of the auditory characterization of the perturbed phoneme combining a shift in mean and reduction in variance. Dashed lines correspond to the perturbed condition. The perturbation goes in the direction of /a/ (horizontal arrow).

https://doi.org/10.1371/journal.pcbi.1005942.g009

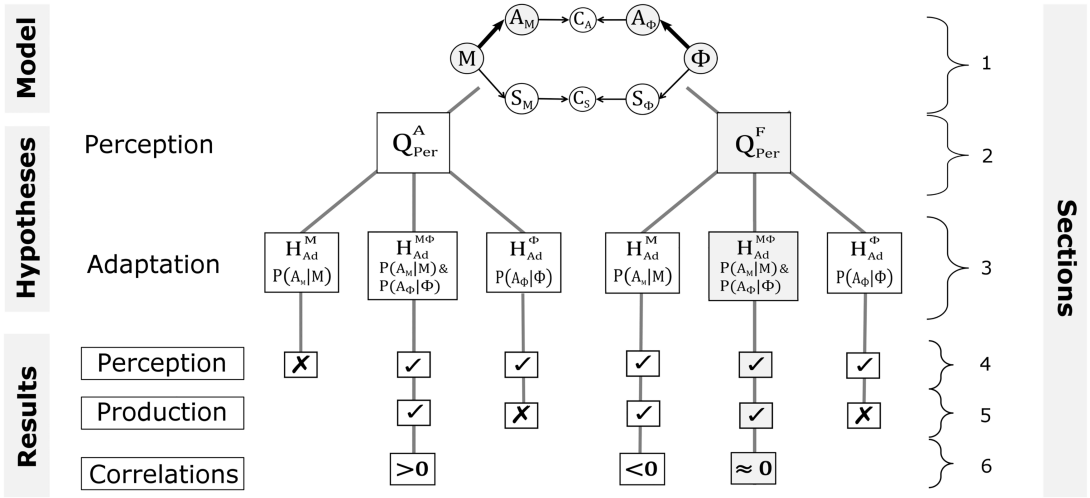

Fig 10. Evaluation of hypotheses about speech perception and adaptation in our modeling work. The model (Section 1 Model definition) enables to define different hypotheses about the sensory pathways involved in speech perception (Section 2 Formulation of speech production and perception questions) and the learning mechanisms involved in adaptation (Section 3 Adaptation hypotheses). Concerning speech perception, we evaluate hypotheses Q^{A}_{Per} and Q^{F}_{Per}, corresponding to the involvement of the auditory pathway only, or the fusion of somatosensory and auditory pathways. Concerning adaptation, we evaluate hypotheses H^{M}_{Ad}, H^{Φ}_{Ad} and $H^{M\Phi}_{\text{Ad}}$, corresponding respectively to an update of the auditor-motor internal model, an update of the auditory characterization of the perturbed phoneme, or both updates simultaneously. Combining hypotheses about perception and adaptation further enables to identify and test different scenarios simulating the experimental paradigm of Lametti et al. [1]. Scenarios leading to perceptual changes incompatible with those observed by Lametti et al. [1] are discarded (x-boxes, Section 4 Evaluation with respect to perception). The remaining scenarios (✓-boxes) are evaluated with respect to their predictions of compensation in production (Section 5 Evaluation with respect to production) and only those that are consistent with results of Lametti et al. [1] are kept. Finally, we evaluate the last scenarios with respect to correlations between perceptual boundary shift and compensation magnitude (Section 6 Evaluation with respect to correlations). The only scenario that matches the no-correlation observation of Lametti et al. [1] is gray-shaded.

https://doi.org/10.1371/journal.pcbi.1005942.g010

hypotheses (Q^{A}_{Per} or Q^{F}_{Per}), reproduce the key-observations described in L-14 concerning the perceptual boundary shifts after adaptation to the auditory perturbation of vowel /ɛ/:

- H^{M}_{Ad}: the auditory-motor internal model is locally updated during adaptation, in the region where the subject articulates speech during the training phase leading to adaptation. No other update occurs.
- H^{Φ}_{Ad}: only the auditory characterization of vowel /ɛ/ is modified and this modification involves a combined update of its location and width.
- $H^{M\Phi}_{\text{Ad}}$: both stored knowledge mentioned in H^{Φ}_{Ad} and H^{M}_{Ad} are simultaneously updated.

Fig 10 illustrates the different stages of our evaluation process. The three selected hypotheses are represented on the third level of Fig 10 from the top. The fourth level represents the outcomes of the evaluation of each hypothesis with respect to perception. The two last levels will be discussed in the next sections.

Evaluation with respect to production

Let us now evaluate the effect of the three previous adaptation hypotheses, H^{M}_{Ad}, H^{Φ}_{Ad} and $H^{M\Phi}_{\text{Ad}}$ with respect to the production question Q^{F}_{Prod}.

Evaluation of H^{M}_{Ad}. Fig 11 presents the outcome of the planning process Q^{F}_{Prod} (right panel) before (solid lines) and after (dashed lines) updating the internal model. We observe a shift of

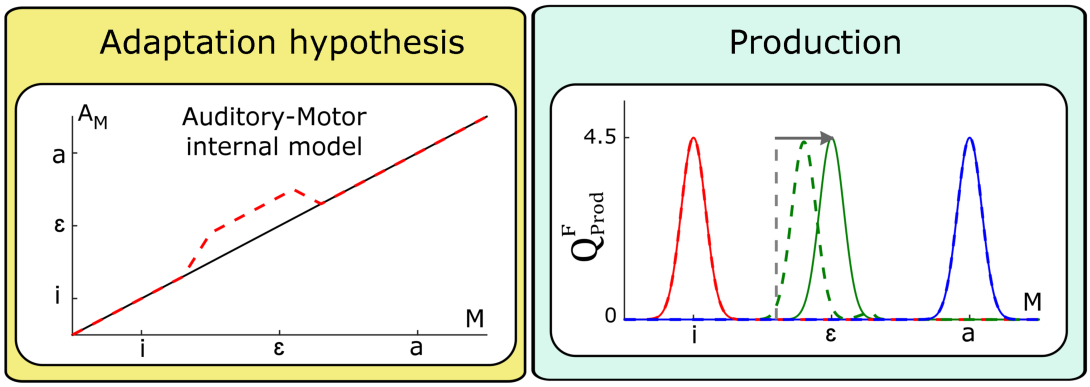

Fig 11. Changes in production question Q^{F}_{Prod} resulting from a local update of the auditory-motor internal model $P(A_M \mid M)$. The update magnitude is equal to the perturbation magnitude (horizontal arrow). The vertical dashed line indicates the shift in control space that would result in complete compensation.

https://doi.org/10.1371/journal.pcbi.1005942.g011

the distribution of the motor commands selected for vowel /ε/ after adaptation, which is in a direction opposite to the perturbation (shift toward /i/, when the auditory perturbation goes toward /a/), in agreement with the reported compensatory behavior.

Consistent with numerous experimental findings [24, 29, 52–54] our model predicts that the compensation for the perturbation of the auditory feedback is not complete. Yet, in the model, the local update of the internal model has been designed in order to enable a full compensation (the magnitude of the change matches the magnitude of the auditory perturbation). However, full compensation does not occur, because the speech planning process takes in consideration both the auditory and the somatosensory characterization of the phoneme. Full compensation would enable a perfect achievement of the auditory characteristics, but at the price of such a large change of the motor commands, that the corresponding somatosensory consequences would not be compatible any longer with the specified somatosensory characterization of the phoneme. Hence, the incomplete compensation is the result of a compromise between the requirements in terms of auditory and somatosensory characteristics. Note as well that the compensatory change in production is restricted to the vowel /ε/. This is a consequence of the locality of the internal model update, which is consistent with experimental observations of transfer of motor learning in speech production [49, 55].

Evaluation of H^{Φ}_{Ad}. Fig 12 presents the outcome of the planning process Q^{F}_{Prod} (right panel) before (solid lines) and after (dashed lines) updating the auditory characterization of vowel /ε/, according to the adaptation hypothesis H^{Φ}_{Ad}. Since, in the context of this hypothesis, all the other representations are unchanged, in particular the auditory-motor internal model, this change of the auditory characterization of vowel /ε/ toward vowel /a/ induces a change of the motor commands that is also toward the articulation of /a/, i.e. in the same direction as the auditory perturbation. This is contrary to the compensatory behavior reported in all the experimental studies involving a perturbation of the auditory feedback.

Evaluation of $H^{M\Phi}_{\text{Ad}}$. Fig 13 presents the outcome of the planning process Q^{F}_{Prod} (right panel) before (solid lines) and after (dashed lines) the combined updates of the auditory-motor internal model and the auditory characterization of vowel /ε/, according to the adaptation hypothesis $H^{M\Phi}_{\text{Ad}}$. We observe that, after these two combined updates, the outcome of the planning process (right panel) is shifted in a direction opposite to the perturbation, in agreement with the reported compensatory behavior. Similarly to the previous evaluation of H^{M}_{Ad}, comparing the magnitude of the shift with the amplitude of the perturbation (horizontal

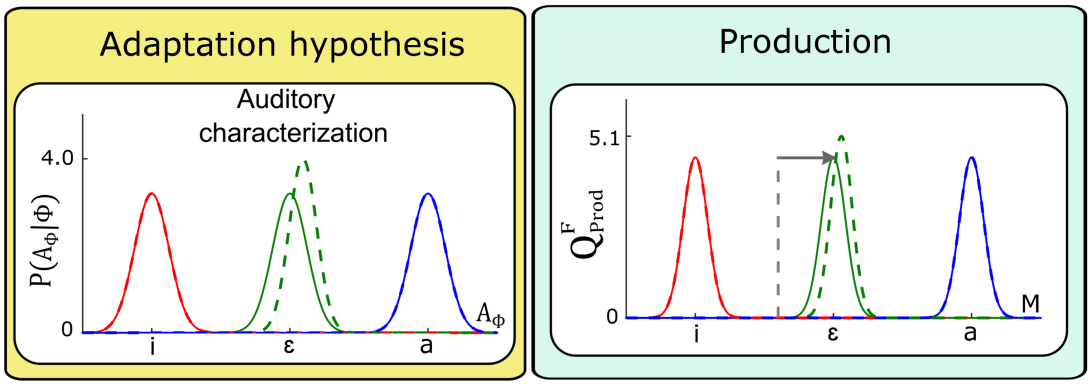

Fig 12. Changes in production question Q^{F}_{Prod} resulting from an update of the auditory characterization of phoneme /ɛ/ $P(A_\Phi \mid [\Phi = /\varepsilon/])$. The update correspond to a combined shift of mean, μ^{ε}_{A} and reduction in standard-deviation σ^{ε}_{A} as previously defined (see Fig 11 for additional details).

https://doi.org/10.1371/journal.pcbi.1005942.g012

arrow) indicates that compensation is not complete. In the present case, the incomplete compensation has a double origin. The shift of the auditory characterization of the perturbed phoneme is an explanation for this phenomenon since this change reduces the need to change articulation in order for the production to match this characterization. In addition, as for H^{M}_{Ad} in Fig 11, compensation is incomplete due to the fact that the new auditory-motor relation leads to auditory and somatosensory states that cannot simultaneously satisfy the two sensory characterizations of phonemes.

Summary. From the three selected adaptation hypotheses, only H^{M}_{Ad} and $H^{M\Phi}_{\text{Ad}}$ are compatible with the compensatory change in production observed in experimental studies. Altogether, as illustrated in Fig 10, we are hence left with three combined perception–adaptation hypotheses that all reproduce the experimental results in perception and production reported in L-14:

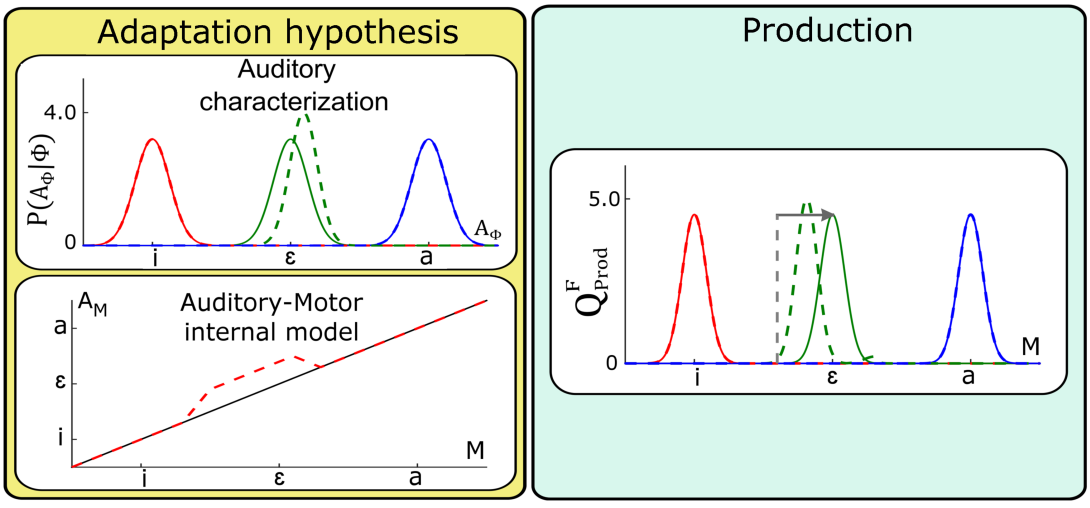

Fig 13. Changes in production question Q^{F}_{Prod} resulting from the combination of the local update of the internal model and the update of the auditory characterization of vowel /ɛ/. See Fig 11 for additional details.

https://doi.org/10.1371/journal.pcbi.1005942.g013

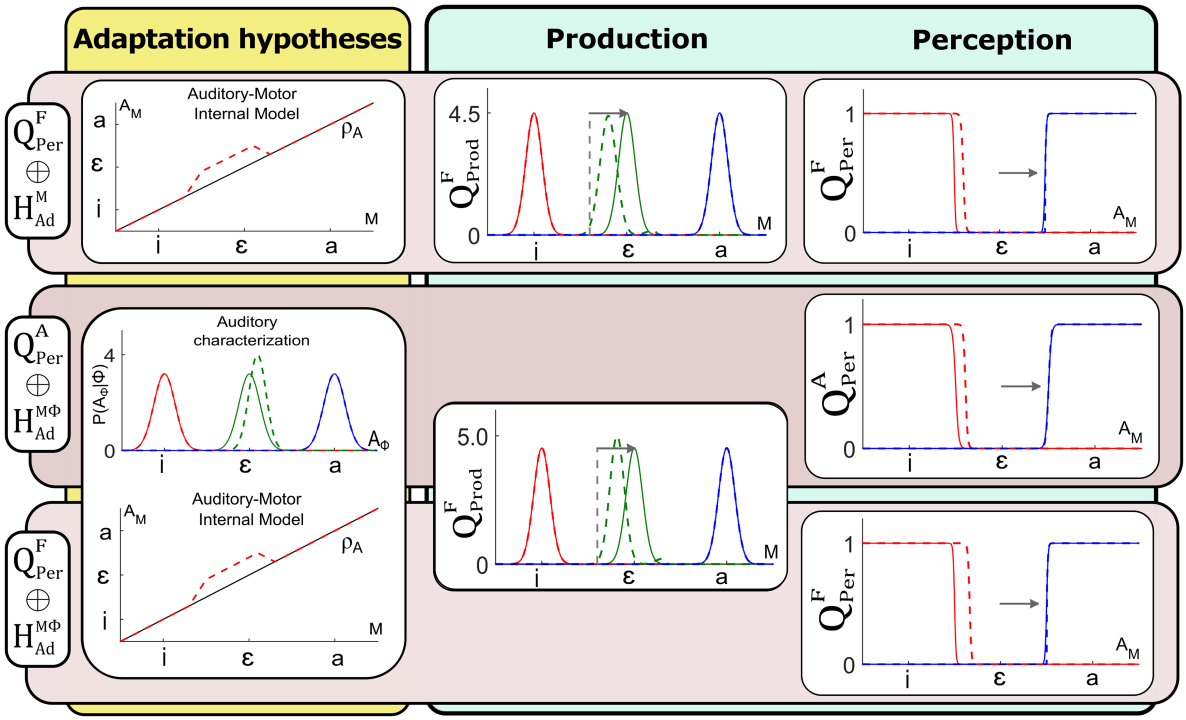

Fig 14. Summary of simulations under each of the three combined hypotheses accounting for the results in L-14.

https://doi.org/10.1371/journal.pcbi.1005942.g014

- $Q^{F}_{\text{Per}} \oplus H^{M}_{\text{Ad}}$: auditory perception is based on the fusion of auditory and somatosensory pathways (Q^{F}_{Per}) and only the auditory-motor internal model is locally updated during adaptation (H^{M}_{Ad}).
- $Q^{A}_{\text{Per}} \oplus H^{M\Phi}_{\text{Ad}}$: auditory perception is based only on the auditory pathway (Q^{A}_{Per}) and both the auditory-motor internal model and the auditory characterization of the perturbed vowel are modified, with a local update for the first and a combined shift and narrowing for the second ($H^{M\Phi}_{\text{Ad}}$).
- $Q^{F}_{\text{Per}} \oplus H^{M\Phi}_{\text{Ad}}$: auditory perception is based on the fusion of sensory pathways (Q^{F}_{Per}) and both the auditory-motor internal model and the auditory characterization of the perturbed vowel are modified, with a local update for the first and a combined shift and narrowing for the second ($H^{M\Phi}_{\text{Ad}}$).

Fig 14 summarizes the corresponding results in production and perception for each of these three selected hypotheses.

As highlighted above in Section "Update of the auditory-motor internal model $P(A_M \mid M)$", the asymmetry of the perceptual boundary shift explained by these three hypotheses is sensitive to the particular values of the parameters involved in the updates according to either hypothesis H^{M}_{Ad} or hypothesis $H^{M\Phi}_{\text{Ad}}$. Other parameter values can lead to boundary changes in both sides of the auditory continuum. The apparent contradiction between studies S-09 and L-14, highlighted in Section "Summary of experimental results we aim at modeling", could thus be interpreted in this context. Refer to Supporting information S5 Text for variations around this theme.

Evaluation with respect to correlations

These three combined hypotheses are equivalent in terms of the qualitative effects predicted with respect to changes in production and perception; they all account for incomplete compensation and for the asymmetric perceptual boundary shift in the direction of perturbation. However, the magnitudes of the perceptual boundary shift and of the motor command shift associated with the compensation differ across the three hypotheses. Experimental studies display large differences across subjects in their capacity to compensate for a perturbation of the auditory feedback [56–58]. Moreover, in L-14 and S-09 subjects differ in the amount of perceptual boundary shift induced by adaptation to the perturbation. If, as suggested in L-14, the perceptual change is mainly due to a change in motor functions, one would expect that subjects who compensate more would exhibit a greater perceptual boundary shift. However, no significant correlation between these two phenomena was found in L-14.

In the present section we focus on this question. First, we identify possible origins for the reported differences concerning the amount of compensation and perceptual shift among subjects. Then, we implement these origins under each of the three combined hypotheses and evaluate their predictions in terms of the correlations between compensation magnitudes and amount of perceptual boundary shift.

Hypotheses on the origins of variability in the magnitude of compensation and perceptual shift. Up to now, we have compared simulations in which for each hypothesis we have arbitrarily chosen a unique set of new parameters for the piece of knowledge that is assumed to be modified during the adaptation process. However, since compensation and adaptation mechanisms in presence of perturbation are highly subject-dependent, we can see our approach as the modeling of a specific subject behavior. In this section we will consider some variations in the changes associated to adaptation in order to investigate the possible consequences of inter-subject variability in the compensation/adaptation process on the categorical boundary shifts in perception.

The adaptation assumptions selected in Section "Effects of combined update hypotheses" involved the local update of the auditory-motor internal model $P(A_M \mid M)$ and the update of the auditory characterization of the perturbed phoneme $P(A_\Phi \mid [\Phi = /\varepsilon/])$. Therefore, inter-subject differences in adaptation can be attributed in the model to different update magnitudes in either of these two terms. These different magnitudes may result from inter-subject differences in learning rates, in novelty or error detection, etc. This leads to the two following hypotheses:

- H^M_{Var}: subjects differ in the magnitude of update of their auditory-motor internal model, $P(A_M \mid M)$, some of them achieving a complete update and some others only a partial update.
- H^Φ_{Var}: subjects differ in the amount of shift of their auditory characterization of the perturbed phoneme, $P(A_\Phi \mid [\Phi = /\varepsilon/])$ (still assuming the relation between mean and variance used in Section "Update of the auditory characterization $P(A_\Phi \mid \Phi)$", i.e such that the perceptual boundary shift is present only on one side of the auditory continuum).

In addition to the two previous hypotheses, we previously noted that, in our model, incomplete compensation resulted from a trade-off between the constraints associated with the auditory and the somatosensory characterizations of the phonemes, which are no longer compatible after adaptation. It is important to point out that the result of this trade-off depends only on the relative strength of the constraint imposed by each sensory pathway. In our previous simulations, both sensory constraints were equivalent (same values of parameters characterizing the Gaussian distributions and linear relation between the two sensory domains), meaning that perturbations to each modality would be equally compensated. However, individual

differences in the amount of compensation to auditory and somatosensory perturbations have been reported in speech production: subjects that adapt more to one sensory perturbation tend to adapt less to the other [59]. This has been suggested as evidence that some subjects may rely more on the auditory modality and others more on the somatosensory modality. Such sensory preferences could originate from individual differences in the sensitivity to each kind of sensory feedback [60], which can be modeled in line with the suggestions of Perkell et al. [52, 61], by differences in the parameters σ_A^ϕ and σ_S^ϕ. Small σ_A^ϕ (resp. σ_S^ϕ) values means that the auditory (resp. somatosensory) characterization of the phoneme is very accurate and that the subject strongly relies on this sensory pathway. Large σ_A^ϕ (resp. σ_S^ϕ) values means either that the sensory characterization is quite inaccurate or that the subject does not rely much on this sensory pathway.

Therefore, we consider a third possible hypothesis concerning the origin of the reported differences in compensation between subjects:

- H_{Var}^σ: subjects differ in the relative precision of their sensory characterizations of phonemes. Some may have greater values of parameter σ_A^ϕ compared to σ_S^ϕ and *vice versa*.

Implementing hypotheses and exploring correlations between the magnitude of compensation and perceptual shift. The three previous hypotheses represent three possible origins of the reported differences in the way subjects adapt to perturbations. These hypotheses are not exclusive and all of them may be involved simultaneously. However, in order to simplify the presentation of the results, we firstly focus on the combined effects of hypotheses H_{Var}^M and H_{Var}^Φ. In other words, we first implement simulations combining different values of update of the auditory-motor internal model (H_{Var}^M), and different values of shift of the auditory characterization of the perturbed phoneme (H_{Var}^Φ). Hence, in this first set of simulation we ignore hypothesis H_{Var}^σ and keep values of σ_A^ϕ and σ_S^ϕ equal.

Fig 15 presents the outcome of simulations for different updates of the auditory-motor internal model and different shifts of the auditory characterization of the perturbed phoneme, for a magnitude of the perturbation representing 40% of the distance between neighboring phonemes and towards phoneme /a/. For the internal model, we specified six update amplitudes in order to enable a compensation varying gradually from 0% to 100% of the magnitude of perturbation (when no influence of other factors reduces compensation) (top left panel of Fig 15). For the auditory characterization of vowel /ɛ/, we implemented six values of parameter σ_A^ε, evenly distributed from the original value, used in the normal condition, to half of this value. The six corresponding values for μ_A^ε (top right panel of Fig 15) were computed from the relation that was already used in Section "Update of the auditory characterization $P(A_\Phi \mid \Phi)$" (as stated in hypothesis H_{Var}^Φ).

Middle panels present the magnitude of compensation and perceptual shifts resulting from the combination of previous updates under the three combined hypotheses, $Q_{\text{Per}}^F \oplus H_{\text{Ad}}^M$ (left panel of Fig 15), $Q_{\text{Per}}^A \oplus H_{\text{Ad}}^{M\Phi}$ (middle panel of Fig 15) and $Q_{\text{Per}}^F \oplus H_{\text{Ad}}^{M\Phi}$ (right panel of Fig 15). Colors correspond to the different magnitudes of internal model update and darkness indicates the amplitude of shift of the auditory characterization of the perturbed phoneme, as indicated by plots in the top panels. X-axis represents the magnitude of compensation in units of the perturbation but in the opposite direction. In other words, value 1 corresponds to a shift in production of the same magnitude but opposite direction of the perturbation (complete compensation), value 0 corresponds to no compensation and value -1 corresponds to a shift in production of the same magnitude and same direction as the perturbation. Y-axis represents the amount of perceptual shift in units of the perturbation and in the same direction.

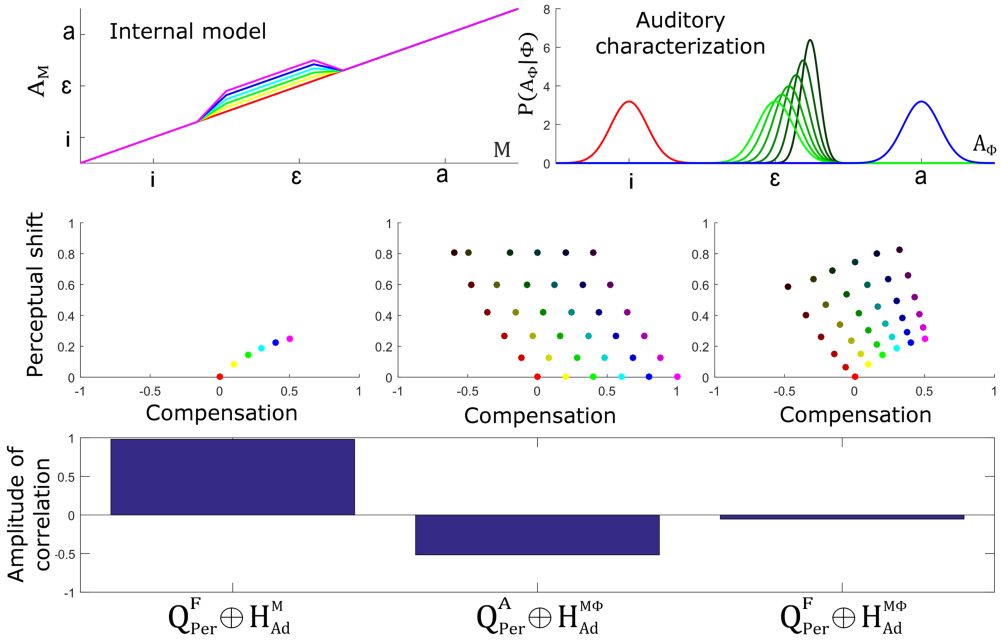

Fig 15. Relation between amplitude of perceptual boundary shift and amount of compensation for hypothesis H^{M}_{Var} and hypothesis H^{Φ}_{Var}. Top panels: considered amplitude of update in the auditory-motor internal model (left) and the amplitude of shift in the auditory characterization of the perturbed phoneme (right). Middle panels: relation between degree of compensation and amount of perceptual shift for each adaptation hypothesis in the case of a local update of the internal model. Colors correspond to the different magnitudes of internal model updates, as indicated in the top left panel. Darkness indicates the amplitude of shift of the auditory characterization of the perturbed phoneme, as indicated in the top right panel. Bottom panel: corresponding amplitude of correlations.

https://doi.org/10.1371/journal.pcbi.1005942.g015

The bottom panel indicates the correlation coefficient between compensation and perceptual shift for the set of data points obtained from the different simulations under each of the three combined hypotheses. Simulations assuming hypothesis $Q^{F}_{Per} \oplus H^{M}_{Ad}$ (left middle panel in Fig 15) show a noticeable positive correlation between magnitude of compensation and perceptual shift. In order to understand this result, it is important to remember that in the context of hypothesis $Q^{F}_{Per} \oplus H^{M}_{Ad}$ only an update of the auditory-motor internal model is assumed. Hence, hypothesis H^{Φ}_{Var} does not apply, and only the effect of inter-subject differences in internal model updates (hypothesis H^{M}_{Var}) can be considered. The magnitude of the articulatory changes associated with compensation is strongly related with the magnitude of the changes provided to the internal model (displacement along the horizontal axis in the figure). Our simulations show that the perceptual boundary shift increases with the magnitude of the changes, but non monotonously: it increases first and becomes stable after. This "saturation" effect is due to the fact that the update of the internal model is local. Altogether, the influence of the update of the internal model in production and perception results in a noticeable positive correlation between the amount of compensation and perceptual shift, contrary to what was reported in L-14.

Simulations assuming hypothesis $Q^{A}_{Per} \oplus H^{M\Phi}_{Ad}$ for the perceptual boundary shift (center panel in Fig 15) show a negative and moderate correlation between amount of compensation and perceptual shift. It should be reminded that in the context of hypothesis $Q^{A}_{Per} \oplus H^{M\Phi}_{Ad}$ adaptation induces both a local update of the motor-auditory internal model and an update of the auditory characterization of vowel /ɛ/, and that perception only involves the auditory pathway. In the absence of any other constraint, the magnitude of the update of the auditory-motor

internal model (hypothesis H^{M}_{Var}) strongly determines the magnitude of the compensation. We have shown above that, when perception only involves the auditory pathway, the update of the auditory-motor internal model has no influence of the perceptual boundary shift. Hence, the update of the internal model does not induce any correlation between the amount of compensation and the magnitude of the perceptual boundary shift. This can be seen in the present simulations where data points corresponding to a given location of the auditory characterization (same darkness) but different values of update of the internal model (different colors) are aligned horizontally.

On the contrary, a shift in the auditory characterization of vowel /ɛ/ has a direct impact on the perceptual boundary shift (positive correlation) and on the amount of compensation (the larger the shift, the smaller the amount of compensation). Thus inter-subject differences in the magnitude of the shift of the auditory characterization of vowel /ɛ/ (hypothesis H^{Φ}_{Var}) result in a negative correlation between the amount of compensation and the magnitude of the perceptual boundary shift. Altogether, in the context of hypothesis $Q^{A}_{\text{Per}} \oplus H^{M\Phi}_{\text{Ad}}$, the combination of hypotheses H^{M}_{Var} and H^{Φ}_{Var} results in a mild negative correlation between the amount of compensation and the magnitude of the perceptual boundary shift, contrary to what was reported in L-14.

Simulations assuming hypothesis $Q^{F}_{\text{Per}} \oplus H^{M\Phi}_{\text{Ad}}$ for the perceptual boundary shift (right middle panel in Fig 15) show an almost vanishing correlation between amount of compensation and perceptual boundary shift. Hypothesis $Q^{F}_{\text{Per}} \oplus H^{M\Phi}_{\text{Ad}}$ can be roughly seen as combining $Q^{F}_{\text{Per}} \oplus H^{M}_{\text{Ad}}$ and $Q^{A}_{\text{Per}} \oplus H^{M\Phi}_{\text{Ad}}$. Since $Q^{F}_{\text{Per}} \oplus H^{M}_{\text{Ad}}$ induces a positive correlation and $Q^{A}_{\text{Per}} \oplus H^{M\Phi}_{\text{Ad}}$ a negative one, the combination of these two influences in $Q^{F}_{\text{Per}} \oplus H^{M\Phi}_{\text{Ad}}$ tends to counterbalance each other, resulting in a much smaller correlation than the two previous ones. For a given shift of the auditory characterization of vowel /ɛ/ (same darkness) simulations with different updates of the internal model (different colors) result in a similar pattern as in the simulations assuming $Q^{F}_{\text{Per}} \oplus H^{M}_{\text{Ad}}$.

However two variations of this pattern can be observed when the shift of the auditory characterization increases. First, the non-linearity induced by the locality of update of the internal model (see above) disappears when the shift increases (darkest *versus* lighter data points). This is due to the fact that the shift in the auditory characterization brings the boundary between phonemes closer to the center of the updated region and reduces the influence of the limits of the local update. (Simulations assuming the general update of the internal model were performed in order to clarify which part of the effects arises from our locality assumption. The obtained results show the same key-properties for both updates of the internal model, which indicates that the obtained pattern of correlations is not an artifact of the particular choice of our local update assumption.)

The second difference is that the slope of the relation between perceptual shift and compensation reduces for greater shifts of the auditory characterization. This is due to the fact that the increase in the magnitude of the shift goes together with a decrease in the width of the auditory characterization of vowel /ɛ/. This results in a stronger influence of the auditory pathway relative to the somatosensory one. Since the influence of the internal model on the perceptual boundary shift is mediated through the somatosensory pathway, the magnitude of the effect reduces when the auditory pathway is stronger. This is consistent with the horizontal alignment obtained under $Q^{A}_{\text{Per}} \oplus H^{M\Phi}_{\text{Ad}}$ where the somatosensory pathway is assumed not to contribute to perception.

In summary, hypothesis $Q^{F}_{\text{Per}} \oplus H^{M\Phi}_{\text{Ad}}$ is more in line with the lack of correlation between compensation and perceptual shift reported in L-14.

We now consider the additional influence of hypothesis H^{σ}_{Var}, assuming variable relative precision of the sensory characterizations of phonemes across subjects. We implemented the

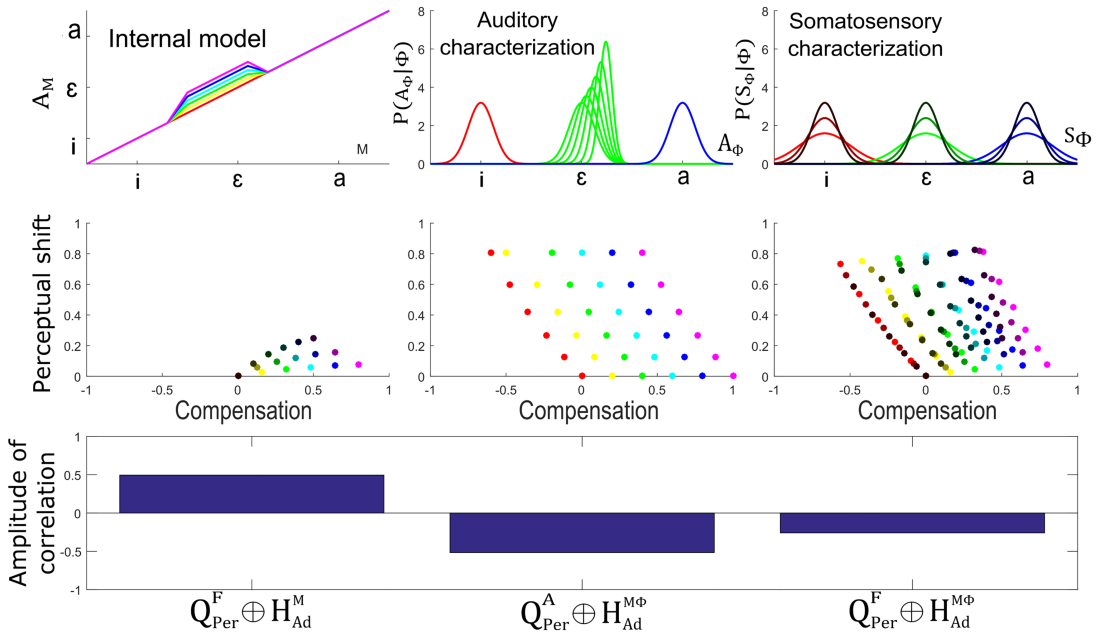

Fig 16. Influence of the combination of hypotheses H^{M}_{Var}, H^{Φ}_{Var} and H^{σ}_{Var} on the relation between compensation and perceptual boundary shift. Changes implemented for the internal model and the auditory characterization are the same as in Fig 15. They are illustrated in the top left and top middle panels. In addition to the previous simulations where parameters of the sensory characterizations, σ^{ϕ}_{S} and σ^{ϕ}_{A}, where both equal, here we implement two additional values of parameters σ^{ϕ}_{S}, corresponding to greater variance of the somatosensory characterization, and therefore reduced weight of the somatosensory pathway. The three values of σ^{ϕ}_{S} implemented are $\frac{1}{8}$ (equal weights of sensory pathways), $\frac{1}{6}$ and $\frac{1}{4}$ (smaller weight for the somatosensory pathway) of the distance between neighboring phonemes. The corresponding changes of the somatosensory characterization are illustrated in the top right panel. Darkness of the colors indicates an increase of the weight of the auditory pathway relatively to the somatosensory pathway.

https://doi.org/10.1371/journal.pcbi.1005942.g016

same simulations as above for different values of parameter σ^{ϕ}_{S}, as illustrated in the top right panel of Fig 16. We retained values of σ^{ϕ}_{S} only greater or equal to the value of σ^{ϕ}_{A} (corresponding to preference on the auditory pathway as compared to the somatosensory pathway), in order to be consistent with the fact that in L-14 only subjects who showed significant compensation to auditory perturbations were kept. Notice that simulations implementing reciprocal values for σ^{ϕ}_{A} and σ^{ϕ}_{S} were also performed. Results are qualitatively similar to those presented below, indicating that they are not a consequence of the particular asymmetric choice implemented here. In Fig 16, the level of darkness of the colors indicates an increasing precision of the somatosensory regions.

Results are consistent with the idea that relative precision of sensory characterizations modulates the influence of each sensory pathway: wider somatosensory characterizations (light colors) are associated with a larger influence of the auditory pathway. As a consequence, in the case of $Q^{F}_{Per} \oplus H^{M}_{Ad}$ (Fig 16, left column) this results in smaller slopes in the relation between perceptual shift and compensation (middle horizontal panel), and decreases the positive correlation between compensation and perceptual shift accordingly (lower horizontal panel). However, for hypothesis $Q^{A}_{Per} \oplus H^{M\Phi}_{Ad}$ (Fig 16, center column) in which the somatosensory pathway is assumed not to contribute to perception, varying somatosensory weight has no consequence on perceptual shift (middle horizontal panel) nor on the resulting correlation (lower horizontal panel). Finally, we observe that hypothesis H^{σ}_{Var} has a small impact on the correlation coefficient obtained assuming $Q^{F}_{Per} \oplus H^{M\Phi}_{Ad}$ (Fig 16, right column) which remains close to zero.

In summary, including hypothesis H^{σ}_{Var} in our simulations, which corresponds to the implementation of individual differences on the weighting of each sensory pathway, confirms that the absence of correlation reported by L-14 may be best attributed to hypothesis $Q^{F}_{\text{Per}} \oplus H^{M\Phi}_{\text{Ad}}$.

Discussion

Using our model, implemented in the Bayesian programming framework, we have been able to implement and test different hypotheses concerning speech motor adaptation to perturbed auditory feedback. In this framework, processes are not directly modeled but are derived from a common set of knowledge, which is represented by means of a joint probability distribution. Hence, in this approach, perception and production processes become naturally related since changes to the underlying knowledge may impact them together. Note that this framework is not restricted to speech, but may be of interest in other areas where production and perception processes have been shown to interact (for instance in the arm motor control literature, see Haith et al. [62] and Ito et al. [63] for alternative approaches, see also Gilet et al. [36] in the context of joint modeling of perception and production of isolated cursive letters).

We have applied this framework to study the perceptual changes that result from motor learning in adaptation to an auditory perturbation in speech. To do so, we have proposed a number of hypotheses about the changes to the common underlying knowledge that may result from motor learning and we have investigated how these changes may give rise to the observed changes in perception and production. This approach has allowed us to identify different possible origins that all may contribute to these changes, supporting but also specifying the interpretation proposed by Lametti et al. [1].

Our experimental simulations provide a number of major results: (1) the induced perceptual shift may actually be compatible with either an auditory or a combined auditory and somatosensory characterization of perceptual targets; (2) the incomplete motor response to auditory perturbations may be due to a mixture of components, related to the combined specification of the phonemic targets for speech production in auditory and somatosensory terms; (3) the asymmetry in perceptual compensation observed in L-14 is also compatible with both theoretical frameworks in speech perception, but actually appears to be sensitive to fine tuning of the experimental parameters in the simulations; (4) patterns of correlations between perceptual and motor responses may be driven by various factors that shed a crucial light on final interpretations of the experimental data.

Of course, these simulations quantitatively depend on a number of modeling choices introduced in Section "Selected aspects for modeling", that are aimed at making simulations tractable and easy to analyze and interpret. This basically includes: (1) the assumption that sensory and motor spaces are one-dimensional, (2) the assumption that sensory-motor mappings are linear (Eqs (2–5, 15 and 16), and (3) the specific tuning of parameters considered in the update hypotheses for adaptation. Still, it is important to stress that the four major results summarized previously have an intrinsic validity, which makes them largely independent of the specific modeling choices. This is due to two major reasons. Firstly, the modeling framework introduced in this work has actually been developed over the years completely independently of the experimental data discussed here. This framework is essentially conceived as a general architecture for formalizing classical assumptions about perceptuo-motor relationships in speech communication [12–14].

Secondly, the four major results appear as general, and likely to be obtained whatever the specific choices in the model. Indeed, the first, second and fourth of these results express direct consequences of the model architecture, in which multisensory fusion (between auditory and

somatosensory representations) in speech production and possibly in speech perception naturally result in trading relationships leading to (1) perceptual adaptation in response to the motor adaptation (2) incomplete response to perturbation and (3) various types of correlation patterns between motor and perceptual adaptation. The case of the third result (asymmetry in perceptual compensation) is quite interesting in this respect. Indeed, it is, contrary to the others, largely ad hoc and related to the specific modeling choices (i.e., the precise relation between mean and variance and parameters of the local update of the internal model, see Supporting information S3 and S4 Text). This makes it fragile and probably not very robust experimentally. But this fragility can also be construed as a prediction: it means that asymmetries should vary from one study to the other, and that this observation is probably not as reliable as what was expected by the authors of L-14 (see for instance a recent study by Schuerman et al. [64] where no significant boundary shift was obtained).

Interestingly, the symmetric vs. asymmetric nature of the perceptuo-motor adaptation process should also largely depend on the nature of the motor-to-sensory internal model, and it is quite well-known that the motor-to-sensory relationship is indeed highly nonlinear, and likely to vary greatly depending on the involved region of the motor or sensory space. This could well explain the difference between the study by Lametti et al. [1] on vowels, that shows a lack of perceptual shift in the region of the auditory space related to what subjects heard in presence of the perturbation, and the study by Shiller et al. [2] in which a perceptual shift in the corresponding regions with fricatives was observed.

Finally, with respect to the one-dimensional assumption, including additional dimensions in sensory and motor spaces may certainly bring interesting behaviors, such as trading relations between dimensions in compensation. However, the /i ɛ a/ continuum considered in L-14 can be basically seen as one-dimensional both in the articulatory space in which the location of the highest point of the tongue is controlled along the high/front—low/back dimension thanks to strong correlations between jaw opening and tongue position [65], —and in the acoustic space with correlated variations between F1 and F2 respectively increasing and decreasing from /i/ to /a/ [66]. Therefore, such additional effects would likely bring only a modulatory change to the magnitude of the resulting shifts in production and perception, without changing the general patterns of results in our simulation.

Therefore, we consider that the simulation results presented here have intrinsic validity. As a consequence, it is of interest to discuss them as some new evidence that can be confronted to important questions related to perceptuo-motor adaptation as discussed in the literature. This is what we will do now, around two points that are the nature of perceptual representations and the origins of incomplete compensation, before introducing some predictions and proposals for new experiments in the field.

Revisiting the interpretation presented in L-14

The first stage of our simulations (Section "Evaluation with respect to perception") both supports and challenges the interpretation by Lametti et al. [1], whereby their data would provide evidence for the role of motor knowledge in speech perception. On the one hand, hypothesis $Q^{F}_{\mathrm{Per}} \oplus H^{M}_{\mathrm{Ad}}$, involving only an update of motor functions, is compatible with their interpretation and in fact also specifies it. Indeed, under this hypothesis a local compensation for the perturbation is required to generate a pattern of perceptual adaptation fitting the asymmetry reported in L-14. On the other hand, in the context of hypothesis $Q^{A}_{\mathrm{Per}} \oplus H^{M\Phi}_{\mathrm{Ad}}$, involving both a local update of the auditory-motor internal model and a modification of the auditory characterization of the perturbed phoneme, a pure auditory theory of speech perception (Q^{A}_{Per}) also provides a pattern of perceptual shifts compatible with their data, even including asymmetries

that were considered as key in their reasoning against auditory theories. In this case, changes in the auditory characterization of a phoneme, involving a coordinated shift of the center of its characterization and a reduction of its variance, are required to explain their results.

It is important to note that it is not unrealistic to assume that motor learning can induce such coordinated changes. Indeed, the shift in location may be explained by a mechanism aligning the auditory characterization of a vowel with its actual realization in presence of the auditory feedback perturbation. The reduction of variance could be attributed to the well-known selective adaptation phenomenon, as suggested by Kleinschmidt et al. [67]: the repeated exposure to the same sound tends to make listeners more sensitive to variations of this sound. Note that, in S-09, selective adaptation was mentioned in order to explain the small perceptual boundary shift observed in their control group after the repeated exposure to the unaltered fricative /s/.

Therefore, at this stage, both an audio-motor and a pure auditory theory may be compatible with the data in L-14. However, the analysis, based on correlations between the amplitude of the perceptual shift and the magnitude of the compensation, indicates that none of the two previous interpretations is compatible with the observations described in L-14. Only hypothesis $Q^{F}_{\mathrm{Per}} \oplus H^{M\Phi}_{\mathrm{Ad}}$, assuming the fusion of sensory pathways in speech perception and adaptation involving the combined updates of the auditory-motor internal model and the auditory characterization of the perturbed phoneme, was compatible with the absence of significant correlation reported in L-14.

In summary, our results support and clarify the initial interpretation of Lametti et al. [1]. By exploiting perceptuo-motor correlations, our results support the claim that both sensory and motor processes intervene in the observed perceptual shift. This result certainly speaks in favor of perceptuo-motor theories of speech perception, though further work should be done in order to better assess the relative contributions of each of these two sets of processes [14].

Three suggested origins for incomplete compensation

Interestingly, in our model, all possible explanations of the link between motor learning and perceptual boundary shift are associated with incomplete compensation for the perturbation, even if the magnitude of the local update of the auditory-motor internal model fully matches the amplitude of the auditory perturbation. This is an important prediction of our model, since incomplete compensations have been systematically observed in all experiments involving a perturbation of the auditory feedback during speech production.

Three mechanisms can indeed be at the origin of incomplete compensation. Firstly, if motor learning induces only an update of the auditory-motor internal model in the context of a bi-modal speech production process, incomplete compensation comes from the interaction between the somatosensory and the auditory specifications of vowels. Secondly, if motor learning also induces a shift and a reduction of variance of the auditory specification of the perturbed phoneme, this provides an additional counter-influence to compensation and the magnitude of the change of the auditory characterization contributes to incomplete compensation. Thirdly, in all cases, if motor learning induces an update of the auditory-motor internal model, the magnitude of this update influences the extent of the compensation: the smaller the update, the more incomplete the compensation.

All these potential explanations of incomplete compensation for perturbations of the auditory feedback have been previously suggested in the literature. In particular, Katseff et al. [68], among other hypotheses, compared the respective influences on the compensation magnitude of a possible interaction between the auditory and the somatosensory feedback *versus* of a possible shift of the auditory region characterizing the pronounced phoneme. They concluded

that behavioral data about compensation for auditory perturbation published in the literature (including those in S-09) are more compatible with an interaction between the two sensory feedbacks.

According to them, in the case of the data in S-09, if the perceptual boundary shift is due to a shift of the auditory characterization of the perturbed phoneme, this latter shift should have the same small amplitude as the former one. Such a small shift of the auditory characterization of the phoneme could not explain the large magnitude of the reduction in compensation.

Our results allow us to qualify their conclusion. Indeed, we have shown that when the shift of the auditory characterization is associated with a reduction of its variance, the magnitude of this shift can be much larger than the magnitude of the perceptual boundary shift. In this case, the shift of the auditory characterization of the perturbed phoneme would perfectly account for the amplitude of the compensation.

Caveats and future directions

At this stage, we have at our disposal a modeling framework to account for the links between production and perception processes. However, the present work focuses on adaptation, by comparing states before and after learning. Investigating the dynamic process occurring during adaptation could provide interesting further insights into the phenomena associated with adaptation. More specifically, the manner with which compensation strategies integrate sensory feedback would inform about the way the sensory-motor characteristics of speech production are updated during the learning phase. For instance, the completeness of compensation appears to be dependent on the amplitude of the perturbation: greater amplitudes of perturbation induce greater sensory errors which appear to result in smaller percentage of total compensation compared to smaller sensory errors. This result seems to be a general property of sensorimotor learning: indeed it has been reported for speech [53, 68], for eye and arm movements [69, 70] and even for bird song [71]. Still, the mechanisms responsible for this decrease in relative adaptation in the case of increasing sensory errors remain unclear.

Our model, in its current state, does not address this question, since it deals only with the consequences of parameters updates, and not with how these updates happen during the learning phase. However, the three possible origins of incomplete compensation (discussed in Section "Three suggested origins for incomplete compensation") actually suggest three possible mechanisms whereby different magnitudes of sensory error would result in different degrees of compensation completeness. First, at the level of the sensory motor mappings, larger sensory errors may drive slower update in order to avoid a faulty reorganization of the learned mapping in the case of totally unexpected and inappropriate sensory signals (see for instance the work of [72] for a modeling approach in line with this idea). Second, at the level of the relative weighting of sensory pathways, the magnitude of sensory errors could disadvantage the pathway with larger errors, assuming that large unexpected errors would arise from inaccurate sensors, which would then be considered unreliable. Finally, at the level of the sensory characterization of the target, larger sensory perturbations may drive larger shifts of the intended target, resulting in smaller amounts of compensation compared to baseline. Each of these hypotheses deserves more careful analysis in light of the existing experimental data: for example, the third hypothesis appears unlikely, since, after the removal of the perturbation, subjects usually return close to the original baseline. Still, these three hypotheses definitely deserve further experimental focus.

Interestingly, our model gives different predictions for these three hypotheses. For instance, if larger sensory errors disadvantage the weighting of one of the sensory pathways, the model would predict that subjects would begin to compensate more for perturbations in

the other sensory modality. Such sensory preferences have been reported previously in speech production [59]; however, to our knowledge, no study has explored the possibility that these preferences may be experimentally modulated by providing larger perturbations to one of the sensory modalities. On the other hand, if sensory errors only influence the update of the sensory-motor mapping or the shift of the sensory characterization of the target, the model would predict no influence of the amount of compensation to perturbations on the other sensory modality. Furthermore, evaluating the influence of the amplitude of perturbation with respect to the resulting perceptual shift could also allow distinguishing between these last two hypotheses. Indeed, if larger sensory errors decrease the update of the sensory-motor mapping, the model would predict a decrease in the amount of perceptual shift, whereas the contrary would happen if larger sensory errors drive greater shifts in the sensory characterization of the target.

Furthermore, as we suggested above, the present model is not limited to the study of auditory perturbations, and investigating the consequences of somatosensory perturbations would allow further evaluation of its pertinence. Indeed, another interesting prediction of the model is that, if adaptation to a somatosensory perturbation updates the somatosensory-motor mapping, it would also induce a boundary shift in the auditory categorization of the perturbed phoneme (but in an opposite direction to perturbation, contrary to the case of auditory perturbations). Such perceptual change following adaptation to a somatosensory perturbation has been actually reported in speech by Nasir and Ostry [3]. Future development of the model would be needed to account for their results, since Nasir and Ostry's paradigm uses a perturbation of the jaw along the horizontal direction, making thus possible a perturbation of the somatosensory feedback without inducing changes in the auditory domain.

More generally, the present model provides a powerful framework for testing hypotheses on the relative roles of auditory and somatosensory representations and processes in perceptual and motor responses to perturbations. Indeed, any means likely to modulate one or the other input (e.g., by exploiting inter-individual variability—or by decreasing the salience of one modality relative to the other, by various techniques such as masking or inhibition of a given channel) should modify the amount of response to perturbations, and thus generate specific quantitative predictions to be compared with new experimental data (e.g., [73]).

Finally, it could be interesting to relate our computational framework with putative neuroanatomical networks suggested by neurocognitive data from the literature. As a matter of fact, a number of studies have explored the neuroanatomy of circuits in charge of monitoring responses to auditory or somatosensory perturbations in speech production (e.g., [74–81]). Even though this is out of the focus of the present study, we have already undertaken studies suggesting possible neuroanatomical correlates of the generic COSMO model [82], which is compatible with the current computational model. A future step in this direction is to adapt the generic architecture to the specific processes associated to perturbation compensation. This would be necessary for better addressing the dynamic adaptation processes mentioned previously in this section.

Conclusions

In order to better understand the mechanisms underlying the observations reported by Lametti et al. [1], we have elaborated a simplified Bayesian model of speech production and speech perception in which phonemes are characterized both in somatosensory and auditory terms. Speech production is assumed to be guided by both sensory characterizations (hypothesis Q^{F}_{Prod}). Two hypotheses concerning speech perception processes were evaluated:

(1) speech perception relies only on the auditory pathway (hypothesis Q^{A}_{Per}), or (2) speech perception relies on the fusion of both auditory and somatosensory pathways (hypothesis Q^{F}_{Per}). We have also considered different hypotheses on the possible consequences of motor adaptation: (1) an update of the auditory-motor internal model, (2) an update of the auditory characterization of the perturbed phoneme, and (3) an update of its somatosensory characterization. Taken separately or in combination, these three update hypotheses lead to seven possible adaptation hypotheses. Combined with the two perception hypotheses Q^{A}_{Per} and Q^{F}_{Per}, these adaptation hypotheses lead to different possible scenarios for explaining the observations of the study of Lametti et al. [1].

In the context of our Bayesian model, we have compared the predictions of these possible scenarios with the experimental observations reported by Lametti et al. [1]. Considering results in perception and production, our simulations indicate that three combined perception-adaptation hypotheses can reproduce the characteristics of the perceptual boundary shift observed in L-14: (1) speech perception relies both on the somatosensory and auditory pathways, and motor adaptation induces only a local update of the auditory-motor internal model ($Q^{F}_{\mathrm{Per}} \oplus H^{M}_{\mathrm{Ad}}$); (2) speech perception relies only on the auditory pathway and motor adaptation induces both a local update of the auditory-motor internal model and the combined shift and size reduction of the auditory characterization of the perturbed phoneme ($Q^{A}_{\mathrm{Per}} \oplus H^{M\Phi}_{\mathrm{Ad}}$), (3) speech perception relies both on the somatosensory and auditory pathways and motor adaptation induces both a local update of the auditory-motor internal model and the combined shift and size reduction of the perturbed phoneme ($Q^{F}_{\mathrm{Per}} \oplus H^{M\Phi}_{\mathrm{Ad}}$).

From that basis, these three selected hypotheses were further evaluated with respect to the predicted correlation between compensation in production and perceptual shift. Our results indicate that only the third hypothesis ($Q^{F}_{\mathrm{Per}} \oplus H^{M\Phi}_{\mathrm{Ad}}$) is able to account for the absence of correlation reported by Lametti et al. [1].

Altogether, this computational approach strengthens and specifies the interpretation by Lametti et al. [1] of their experimental data in favor of perceptuo-motor links in speech perception. Our model provides novel insights into the mechanisms influencing speech perception and production after adaptation to perturbations of the auditory feedback. Future work should focus on the dynamics of adaptation as well as on the relation between the degree of adaptation and the amount of perceptual changes.

Supporting information

S1 Text. Detailed model definition.
(PDF)

S2 Text. Derivation of Bayesian inference equations.
(PDF)

S3 Text. Specification of parameters for the local update of the auditory-motor mapping ρ_A.
(PDF)

S4 Text. Specification of parameters of the sensory characterizations of phonemes $P(A_\Phi \mid \Phi)$ and $P(S_\Phi \mid \Phi)$.
(PDF)

S5 Text. From L-14 to S-09: Variations around the theme.
(PDF)

Author Contributions

Conceptualization: Jean-François Patri, Pascal Perrier, Jean-Luc Schwartz, Julien Diard.

Funding acquisition: Jean-Luc Schwartz.

Methodology: Jean-François Patri, Pascal Perrier, Jean-Luc Schwartz, Julien Diard.

Software: Jean-François Patri.

Writing – original draft: Jean-François Patri, Pascal Perrier, Jean-Luc Schwartz, Julien Diard.

Writing – review & editing: Jean-François Patri, Pascal Perrier, Jean-Luc Schwartz, Julien Diard.

References

1. Lametti DR, Rochet-Capellan A, Neufeld E, Shiller DM, Ostry DJ. Plasticity in the Human Speech Motor System Drives Changes in Speech Perception. Journal of Neuroscience. 2014; 34(31):10339–10346. https://doi.org/10.1523/JNEUROSCI.0108-14.2014 PMID: 25080594
2. Shiller DM, Sato M, Gracco VL, Baum SR. Perceptual recalibration of speech sounds following speech motor learning. The Journal of the Acoustical Society of America. 2009; 125(2):1103–1113. https://doi.org/10.1121/1.3058638 PMID: 19206885
3. Nasir SM, Ostry DJ. Auditory plasticity and speech motor learning. Proceedings of the National Academy of Sciences. 2009; 106(48):20470–20475. https://doi.org/10.1073/pnas.0907032106
4. Blumstein SE, Stevens KN. Phonetic features and acoustic invariance in speech. Cognition. 1981; 10 (1-3):25–32. https://doi.org/10.1016/0010-0277(81)90021-4 PMID: 7198546
5. Liberman AM, Mattingly IG. The motor theory of speech perception revised. Cognition. 1985; 21(1):1–36. https://doi.org/10.1016/0010-0277(85)90021-6 PMID: 4075760
6. Stevens KN. On the quantal nature of speech. Journal of Phonetics. 1989; 17:3–45.
7. Fowler CA. Listeners do hear sounds, not tongues. The Journal of the Acoustical Society of America. 1996; 99(3):1730–1741. https://doi.org/10.1121/1.415237 PMID: 8819862
8. Schwartz JL, Basirat A, Ménard L, Sato M. The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. Journal of Neurolinguistics. 2012; 25(5):336–354. https://doi.org/10.1016/j.jneuroling.2009.12.004
9. Patri JF, Diard J, Perrier P. Modélisation bayésienne de la planification motrice des gestes de parole: Évaluation du rôle des différentes modalités sensorielles. In: J.E.P. 2016. vol. 1; 2016. p. 419–427.
10. Cressman EK, Henriques DYP. Sensory recalibration of hand position following visuomotor adaptation. Journal of Neurophysiology. 2009; 102(6):3505–3518. https://doi.org/10.1152/jn.00514.2009 PMID: 19828727
11. Ostry DJ, Darainy M, Mattar AAG, Wong J, Gribble PL. Somatosensory plasticity and motor learning. The Journal of Neuroscience. 2010; 30(15):5384–93. https://doi.org/10.1523/JNEUROSCI.4571-09.2010 PMID: 20392960
12. Moulin-Frier C, Laurent R, Bessière P, Schwartz JL, Diard J. Adverse conditions improve distinguishability of auditory, motor and perceptuo-motor theories of speech perception: an exploratory Bayesian modeling study. Language and Cognitive Processes. 2012; 27(7–8):1240–1263. https://doi.org/10.1080/01690965.2011.645313
13. Moulin-Frier C, Diard J, Schwartz JL, Bessière P. 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"). 2015; 53:5–41.
14. Laurent R, Barnaud ML, Schwartz JL, Bessière P, Diard J. The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception. Psychological Review. 2017;. https://doi.org/10.1037/rev0000069 PMID: 28471206
15. Diard J. Bayesian Algorithmic Modeling in Cognitive Science [Habilitation à Diriger des Recherches (HDR)]. Université Grenoble Alpes; 2015.
16. Bessière P, Mazer E, Ahuactzin JM, Mekhnacha K. Bayesian Programming. Boca Raton, Florida: CRC Press; 2013.
17. Marr D. Vision. A Computational Investigation into the Human Representation and Processing of Visual Information. New York, USA: W.H. Freeman and Company; 1982.

18. Barnaud ML, Schwartz JL, Diard J, Bessière P. Sensorimotor learning in a Bayesian computational model of speech communication. In: The Sixth Joint IEEE International Conference Developmental Learning and Epigenetic Robotics (ICDL-EPIROB 2016). IEEE; 2016. p. 27–32.
19. Barnaud ML, Diard J, Bessière P. Assessing idiosyncrasies in a Bayesian model of speech communication. In: Interspeech 2016; 2016. p. 2080–2084.
20. Patri JF, Diard J, Perrier P. Optimal speech motor control and token-to-token variability: a Bayesian modeling approach. Biological Cybernetics. 2015; 109(6):611–626. https://doi.org/10.1007/s00422-015-0664-4 PMID: 26497359
21. Patri JF, Perrier P, Diard J. Bayesian modeling in speech motor control: a principled structure for the integration of various constraints. In: Interspeech 2016. San Francisco; 2016. p. 3588–3592.
22. Sanguineti V, Laboissière R, Ostry DJ. A dynamic biomechanical model for neural control of speech production. The Journal of the Acoustical Society of America. 1998; 103(3):1615–1627. https://doi.org/10.1121/1.421296 PMID: 9514026
23. Atal BS, Chang JJ, Mathews MV, Tukey JW. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. The Journal of the Acoustical Society of America. 1978; 63 (5):1535–1555. https://doi.org/10.1121/1.381848 PMID: 690333
24. Cai S, Ghosh SS, Guenther FH, Perkell JS. Focal Manipulations of Formant Trajectories Reveal a Role of Auditory Feedback in the Online Control of Both Within-Syllable and Between-Syllable Speech Timing. Journal of Neuroscience. 2011; 31(45):16483–16490. https://doi.org/10.1523/JNEUROSCI.3653-11.2011 PMID: 22072698
25. Purcell DW, Munhall KG. Compensation following real-time manipulation of formants in isolated vowels. The Journal of the Acoustical Society of America. 2006; 119(4):2288–2297. https://doi.org/10.1121/1.2173514 PMID: 16642842
26. Tremblay S, Shiller DM, Ostry DJ. Somatosensory basis of speech production. Nature. 2003; 423 (6942):866–869. https://doi.org/10.1038/nature01710 PMID: 12815431
27. Nasir SM, Ostry DJ. Speech motor learning in profoundly deaf adults. Nature Neuroscience. 2008; 11 (10):1217–22. https://doi.org/10.1038/nn.2193 PMID: 18794839
28. Savariaux C, Perrier P, Orliaguet JP. Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: a study of the control space in speech production. The Journal of the Acoustical Society of America. 1995; 98(5):2428–2442. https://doi.org/10.1121/1.413277
29. Houde JF, Jordan MI. Sensorimotor adaptation of speech I: compensation and adaptation. Journal of Speech, Language, and Hearing Research. 2002; 45(2):295–310. https://doi.org/10.1044/1092-4388(2002/023) PMID: 12003512
30. Scott SK, Wise RJS. The functional neuroanatomy of prelexical processing in speech perception. Cognition. 2004; 92(1-2):13–45. https://doi.org/10.1016/j.cognition.2002.12.002 PMID: 15037125
31. Hickok G, Poeppel D. The cortical organization of speech processing. Nature Reviews Neuroscience. 2007; 8(5):393–402. https://doi.org/10.1038/nrn2113 PMID: 17431404
32. Rauschecker JP, Scott SK. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Neuroscience. 2009; 12(6):718–724. https://doi.org/10.1038/nn.2331 PMID: 19471271
33. Friederici AD. The cortical language circuit: from auditory perception to sentence comprehension. Trends in Cognitive Sciences. 2012; 16(5):262–268. https://doi.org/10.1016/j.tics.2012.04.001 PMID: 22516238
34. Ito T, Tiede M, Ostry DJ. Somatosensory function in speech perception. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106(4):1245–8. https://doi.org/10.1073/pnas.0810063106 PMID: 19164569
35. Skipper JI, Devlin JT, Lametti DR. The hearing ear is always found close to the speaking tongue: review of the role of the motor system in speech perception. Brain and Language. 2017; 164:77–105. https://doi.org/10.1016/j.bandl.2016.10.004 PMID: 27821280
36. Gilet E, Diard J, Bessière P. Bayesian action—perception computational model: interaction of production and recognition of cursive letters. PLOS ONE. 2011; 6(6):e20387. https://doi.org/10.1371/journal.pone.0020387 PMID: 21674043
37. Kawato M. Internal models for motor control and trajectory planning. Current Opinion in Neurobiology. 1999; 9(6):718–727. https://doi.org/10.1016/S0959-4388(99)00028-8 PMID: 10607637
38. Wolpert DM, Miall RC, Kawato M. Internal models in the cerebellum. Trends in Cognitive Sciences. 1998; 2(9):338–347. https://doi.org/10.1016/S1364-6613(98)01221-2 PMID: 21227230
39. Jordan MI, Rumelhart DE. Forward models: supervised learning with a distal teacher. Cognitive Science. 1992; 16(3):307–354. https://doi.org/10.1207/s15516709cog1603_1

40. Ostry DJ, Feldman AG. A critical evaluation of the force control hypothesis in motor control. Experimental Brain Research. 2003; 153(3):275–288. https://doi.org/10.1007/s00221-003-1624-0 PMID: 14610628

41. Pilon JF, De Serres SJ, Feldman AG. Threshold position control of arm movement with anticipatory increase in grip force. Experimental Brain Research. 2007; 181(1):49–67. https://doi.org/10.1007/s00221-007-0901-8 PMID: 17340124

42. Kleinschmidt DF, Jaeger TF. Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. Psychological Review. 2015; 122(2):148–203. https://doi.org/10.1037/a0038695 PMID: 25844873

43. De Boer B, Kuhl PK. Investigating the role of infant-directed speech with a computer model. Acoustics Research Letters Online. 2003; 4(4):129–134. https://doi.org/10.1121/1.1613311

44. Clayards M, Tanenhaus MK, Aslin RN, Jacobs RA. Perception of speech reflects optimal use of probabilistic speech cues. Cognition. 2008; 108:804–809. https://doi.org/10.1016/j.cognition.2008.04.004 PMID: 18582855

45. Clayards M, Aslin RN, Tanenhaus MK, Jacobs RA. Within category phonetic variability affects perceptual uncertainty. In: Proc. 16th International Congress of Phonetic Sciences, Saarbrücken, Germany; 2007. p. 701–704.

46. Feldman NH, Griffiths TL, Morgan JL. The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. Psychological Review. 2009; 116(4):752–782. https://doi.org/10.1037/a0017196 PMID: 19839683

47. Fowler CA. An event approach to the study of speech perception from a direct-realist perspective. Status Report on Speech Research, edited by IG Mattingly and N O'Brien, Haskins Laboratories, New Haven, CT. 1986; p. 139–169.

48. Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. Nature. 2002; 415(6870):429–433. https://doi.org/10.1038/415429a PMID: 11807554

49. Tremblay S, Houle G, Ostry DJ. Specificity of Speech Motor Learning. The Journal of Neuroscience. 2008; 28(10):2426–2434. https://doi.org/10.1523/JNEUROSCI.4196-07.2008 PMID: 18322088

50. Rochet-Capellan A, Richer L, Ostry DJ. Nonhomogeneous transfer reveals specificity in speech motor learning. Journal of Neurophysiolgy. 2012; 107(6):1711–1717. https://doi.org/10.1152/jn.00773.2011

51. Mattar AAG, Ostry DJ. Modifiability of generalization in dynamics learning. Journal of Neurophysiology. 2007; 98:3321–3329. https://doi.org/10.1152/jn.00576.2007 PMID: 17928561

52. Villacorta VM, Perkell JS, Guenther FH. Sensorimotor adaptation to perturbations of vowel acoustics and its relation to perception. The Journal of the Acoustical Society of America. 2007; 122(4):2306–2319. https://doi.org/10.1121/1.2773966 PMID: 17902866

53. MacDonald EN, Goldberg R, Munhall KG. Compensations in response to real-time formant perturbations of different magnitudes. The Journal of the Acoustical Society of America. 2010; 127(2):1059–1068. https://doi.org/10.1121/1.3278606 PMID: 20136227

54. Caudrelier T, Perrier P, Schwartz JL, Rochet-Capellan A. Does auditory-motor learning of speech transfer from the CV syllable to the CVCV word? In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. vol. 08-12-Sept; 2016. p. 2095–2099.

55. Rochet-Capellan A, Ostry DJ. Simultaneous acquisition of multiple auditory-motor transformations in speech. The Journal of Neuroscience: the official journal of the Society for Neuroscience. 2011; 31(7):2657–2662. https://doi.org/10.1523/JNEUROSCI.6020-10.2011

56. Houde JF, Jordan MI. Sensorimotor Adaptation in Speech Production. Science (New York, NY). 1998; 1(5354):1213–1216. https://doi.org/10.1126/science.279.5354.1213

57. Cai S, Ghosh SS, Guenther FH, Perkell JS. Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong /iau/ and its pattern of generalization. The Journal of the Acoustical Society of America. 2010; 128(4):2033–48. https://doi.org/10.1121/1.3479539 PMID: 20968374

58. MacDonald EN, Purcell DW, Munhall KG. Probing the independence of formant control using altered auditory feedback. The Journal of the Acoustical Society of America. 2011; 129(2):955–965. https://doi.org/10.1121/1.3531932 PMID: 21361452

59. Lametti DR, Nasir SM, Ostry DJ. Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. The Journal of Neuroscience: the official journal of the Society for Neuroscience. 2012; 32(27):9351–9358. https://doi.org/10.1523/JNEUROSCI.0404-12.2012

60. Yates AJ. Delayed auditory feedback and shadowing. The Quarterly Journal of Experimental Psychology. 1965; 17(2):125–131. https://doi.org/10.1080/17470216508416421

61. Perkell JS, Lane H, Ghosh S, Matthies ML. Mechanisms of vowel production: auditory goals and speaker acuity. In: Proceedings of the Eighth International Seminar on speech production, Strasbourg, France; 2008. p. 29–32.
62. Haith A, Jackson CP, Miall RC, Vijayakumar S. Unifying the sensory and motor components of sensorimotor adaptation. In: Advances in Neural Information Processing Systems; 2009. p. 593–600.
63. Ito S, Darainy M, Sasaki M, Ostry DJ. Computational model of motor learning and perceptual change. Biological Cybernetics. 2013; 107(6):653–667. https://doi.org/10.1007/s00422-013-0565-3 PMID: 23989535
64. Schuerman WL, Meyer AS, McQueen JM. Mapping the speech code: cortical responses linking the perception and production of vowels. Frontiers in Human Neuroscience. 2017; 11. https://doi.org/10.3389/fnhum.2017.00161 PMID: 28439232
65. Lindau M. Vowel features. Language. 1978; p. 541–563.
66. Fant G. Feature analysis of Swedish vowels–a revisit. KTH, STL-QPSR. 1983; p. 2–3.
67. Kleinschmidt D, Jaeger TF. A Bayesian belief updating model of phonetic recalibration and selective adaptation. In: 2nd Workshop on Cognitive Modeling and Computational Linguistics. June; 2011. p. 10–19.
68. Katseff S, Houde J, Johnson K. Partial compensation for altered auditory feedback: a tradeoff with somatosensory feedback? Language and Speech. 2012; 55(2):295–308. https://doi.org/10.1177/0023830911417802 PMID: 22783636
69. Robinson FR, Noto CT, Bevans SE. Effect of visual error size on saccade adaptation in monkey. Journal of Neurophysiology. 2003; 90(2):1235–44. https://doi.org/10.1152/jn.00656.2002 PMID: 12711711
70. Wei K, Körding K. Relevance of error: what drives motor adaptation? Journal of Neurophysiology. 2009; 101:655–664. https://doi.org/10.1152/jn.90545.2008 PMID: 19019979
71. Sober SJ, Brainard MS. Vocal learning is constrained by the statistics of sensorimotor experience. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109(51):21099–103. https://doi.org/10.1073/pnas.1213622109 PMID: 23213223
72. Hahnloser RHR, Narula G. A Bayesian account of vocal adaptation to pitch-shifted auditory feedback. PlOS ONE. 2016; p. 1–13.
73. Schuerman WL, Nagarajan S, McQueen JM, Houde J. Sensorimotor adaptation affects perceptual compensation for coarticulation. The Journal of the Acoustical Society of America. 2017; 141(4):2693–2704. https://doi.org/10.1121/1.4979791 PMID: 28464681
74. Houde JF, Nagarajan SS, Sekihara K, Merzenich MM. Modulation of the auditory cortex during speech: an MEG study. Journal of Cognitive Neuroscience. 2002; 14(8):1125–1138. https://doi.org/10.1162/089892902760807140 PMID: 12495520
75. Golfinopoulos E, Tourville JA, Bohland JW, Ghosh SS, Nieto-Castanon A, Guenther FH. fMRI investigation of unexpected somatosensory feedback perturbation during speech. NeuroImage. 2011; 55(3):1324–1338. https://doi.org/10.1016/j.neuroimage.2010.12.065 PMID: 21195191
76. Eliades SJ, Wang X. Neural substrates of vocalization feedback monitoring in primate auditory cortex. Nature. 2008; 453(7198):1102–1106. https://doi.org/10.1038/nature06910 PMID: 18454135
77. Christoffels IK, Formisano E, Schiller NO. Neural correlates of verbal feedback processing: an fMRI study employing overt speech. Human Brain Mapping. 2007; 28(9):868–879. https://doi.org/10.1002/hbm.20315 PMID: 17266104
78. Hashimoto Y, Sakai KL. Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: an fMRI study. Human Brain Mapping. 2003; 20(1):22–28. https://doi.org/10.1002/hbm.10119 PMID: 12953303
79. Fu CHY, Vythelingum GN, Brammer MJ, Williams SCR, Amaro E, Andrew CM, et al. An fMRI study of verbal self-monitoring: neural correlates of auditory verbal feedback. Cerebral Cortex. 2006; 16(7):969–977. https://doi.org/10.1093/cercor/bhj039 PMID: 16195470
80. Tourville JA, Reilly KJ, Guenther FH. Neural mechanisms underlying auditory feedback control of speech. NeuroImage. 2008; 39(3):1429–1443. https://doi.org/10.1016/j.neuroimage.2007.09.054 PMID: 18035557
81. Zheng ZZ, Munhall KG, Johnsrude IS. Functional overlap between regions involved in speech perception and in monitoring one's own voice during speech production. Journal of Cognitive Neuroscience. 2010; 22(8):1770–1781. https://doi.org/10.1162/jocn.2009.21324 PMID: 19642886
82. Barnaud ML, Diard J, Bessière P, Schwartz JL. Computational simulations of perceptuo-motor idiosyncrasies support the involvement of motor knowledge in speech perception. Brain and Language. Forthcoming.

Chapter 11

Conclusion

1 Summary and interest of this work

The main contribution of this thesis has been to apply a Bayesian modeling framework for the study of speech motor planning. Accounting for the token-to-token variability of speech productions of any given speech item has been our starting point, motivating the use of the Bayesian framework. In the Bayesian framework knowledge and uncertainty are formally represented with probability distributions, which can be further combined and manipulated in order to derive new knowledge in the form of Bayesian inference. This enabled us to addressed token-to-token variability in two main steps. Firstly, we formally represented the abundance of possible productions of any given speech item as uncertainty in the knowledge characterizing speech motor goals. This enabled us to formulate speech motor planning as an inference process associating each phoneme with a specific probability distribution in the space of the control variable. Secondly, we have suggested that token-to-token variability results from a decision process that samples the space of the control variable according to the probability distributions obtained during motor planning.

We have implemented this idea by reformulating in the Bayesian framework the motor planning stage of GEPPETO, an existing model of speech motor control developed in our lab. This was our first contribution. We have shown that the Bayesian approach enables to implement the basic hypotheses of GEPPETO, in a more general and mathematically uniform framework, while avoiding the constraints and limitations of optimal methods. More specifically, our first contribution exploits and highlights a particular feature of the Bayesian modeling approach, which enables to naturally address ill-posed problems, such as those that result from the indeterminacy and abundance of solutions in speech motor planning (Colas et al., 2010), without involving regularizations based on cost minimization.

A second interesting feature of the proposed framework is the fact that it offers a structured methodology for the definition of Bayesian models. Our second contribution further exploits and illustrates this feature by extending the B-GEPPETO model to include auditory and somatosensory characterizations of speech motor goals. We have applied and evaluated this multisensory model in the context of well documented adaptation phenomena to auditory and somatosensory perturbations of speech production. In particular, we have explored different approaches for modulating the involvement of each of these two sensory pathways in speech motor planning process, providing two different but equivalent implementations of sensory preferences in the model.

The development of the multisensory extension of B-GEPPETO crucially relies on the existence of a somatosensory characterization of speech motor goals. In line with other models of the literature, we have assumed that these somatosensory goals would emerge from the auditory ones during the speech acquisition process. We have further assumed that somatosensory goals would be learned as simplified local approximations of the somatosensory correlates of

the original auditory target regions of GEPPETO, where by local we mean "not covering the whole range of variations" associated with auditory target regions. Consequently, from this assumption we expected that in situations where only somatosensory feedback would be available, utterances would be less variable and more prototypical than in situations where both auditory and somatosensory feedbacks are available. We further expected that a somatosensory characterization of phonemes would enable an identification of vowels based only on somatosensory information. We presented two experimental studies that aimed at exploring these two specific predictions.

Our third and last contribution exploits another interesting feature of the proposed framework, which offers a principled and unified structure for describing knowledge involved both in perception and production processes. We applied a simplified one-dimensional linear implementation of the multisensory model in order to formulate and evaluate different hypotheses concerning speech motor adaptation and interpret the perceptual changes induced by motor learning under altered auditory feedback as reported in recent experimental studies.

2 Future directions

We would like to conclude this manuscript by discussing a certain number of additional questions raised but not addressed by our work. In Section 2.1, other possible architectures of sensory integration in the model will be discussed. Then, in Section 2.2, caveats will be considered together with additional features that could be further developed and explored with the current framework.

2.1 Alternative architectures of sensory integration

One of the main goals of modeling is to formulate and implement assumption in a precise mathematical framework in order to evaluate them and specify further experimental questions. The contributions of this thesis already treated several such assumptions; however, some alternative exploratory approaches exist that we wish to highlight and discuss. We already formulated one of these questions in Chapter 6 where two alternative architectures of speech motor planning were shown to predict different patterns of variability depending on whether categorical perception was involved in speech motor planning or not.

A similar situation exists for the multisensory model presented in Chapter 7. We defined the multisensory extension of GEPPETO by combining sensory information with the use of coherence variables. In fact, two additional models can be formulated involving the same auditory and somatosensory knowledge than in the model presented in Chapter 7. These alternative formulations, illustrated in Fig 11.1, raise questions about how sensory variables are combined, integrated and whether categorization [7] processes are or not involved at a certain stage.

The model presented in Chapter 7, combined sensory information without involving categorization processes. The predicted sensory consequences of motor variables were directly evaluated with respect to the phoneme-specific expected probability distributions of sensory values (top panel of Fig 11.1). The two considered variants involve sensory categorization in two different ways. The first model (bottom left panel in Fig 11.1) processes the predicted sensory consequences of motor variables separately involving two categorization processes, one auditory ($P(\Phi_A \mid A)$) and one somatosensory ($P(\Phi_S \mid S)$). Categories are then combined with a

[7] By categorization we mean the processing of sensory variables to extract categories.

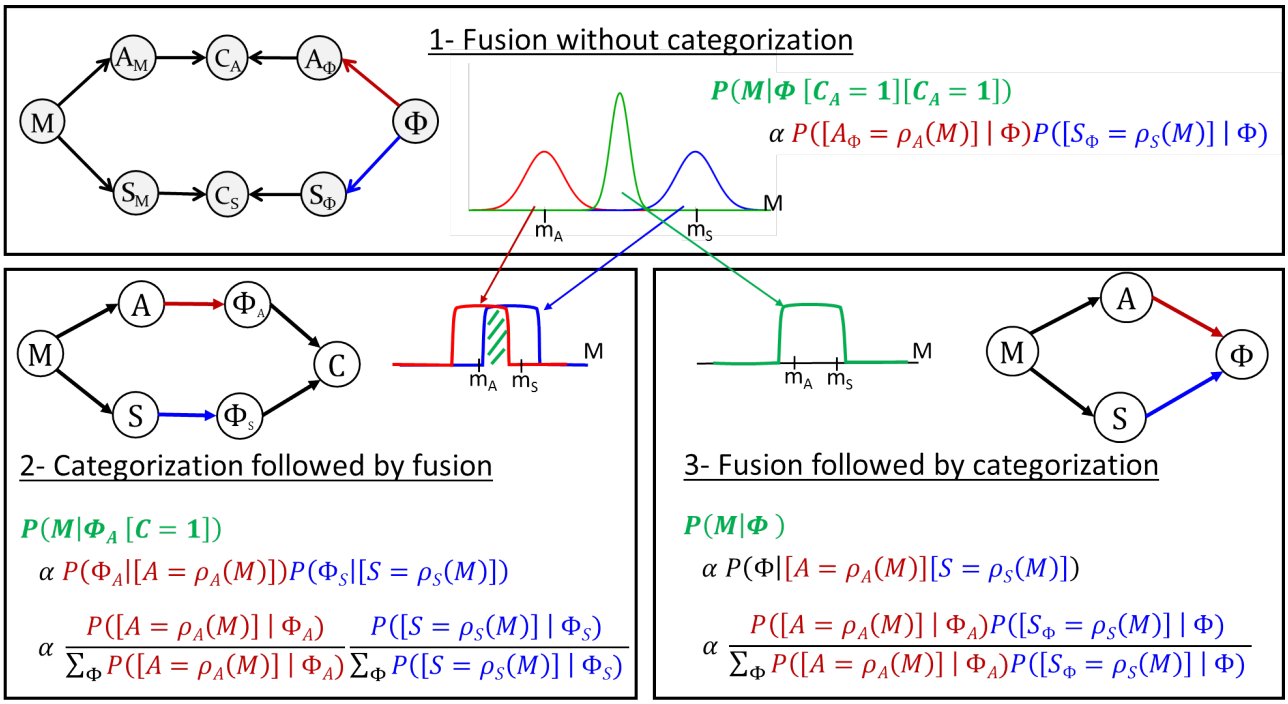

Figure 11.1: Three architectures for sensory integration. *The top panel illustrates the architecture used in the present work and presented in Chapter 7, corresponding to sensory fusion without involvement of a categorization sub-module. The two bottom panels illustrate two alternative architectures involving two categorization modules that are then combined (left panel) or a single categorization module based on the fusion of sensory integration (right panel). Equations corresponding to each inference of motor planning are provided along with diagrams illustrating their respective predictions when auditory and somatosensory pathway do not agree (e.g., in the context of Chapter 8, after adaptation to sensory perturbations). The effort constraint is not considered for simplicity.*

coherence variable. The second model (bottom right panel in Fig 11.1) processes the predicted sensory consequences of motor variables together with respect to a single categorization process ($P(\Phi \mid A\ S)$) that integrates both auditory and somatosensory information.

In Chapter 6, we have discussed how expressing the target regions either directly, by their Gaussian distribution as in model 1 (Top Panel in Fig 11.1), or through inversion, by the result of a categorization process as in models 2 and 3 (Bottom Panels in Fig 11.1), would predict different distributions of produced sounds. This is also the case when target regions are both auditory and somatosensory, with an additional complexity due to the fact that categorization can either proceed independently on auditory and somatosensory spaces, and be combined later, or it can proceed in the joint auditory and somatosensory spaces. These two variants are related to the issue of early vs late fusion process, or of weak vs strong fusion (Landy, Maloney, Johnston, & Young, 1995; Robert-Ribes, 1995; Schwartz, Robert-Ribes, & Escudier, 1998). Our choice on the first of these three models was mainly motivated by mathematical arguments, and more specifically by the property of the coherence variables that enable to connect or disconnect pathways and thus compare different planning processes within one single model architecture. It would be interesting to further explore and evaluate these variants with respect to experimental evidence in order to assess which one better characterizes the architecture of

sensory integration involved in speech motor planning.

The first criterion that would enable to distinguish between the model used in this thesis and the two other variants is the one considered in Chapter 6: because they involve categorical perception modules, the two variants predict plateau distributions in which all control variables are equally relevant for the correct production of the intended phoneme (bottom panels in of Fig 11.1), whereas the model used in this thesis which does not involve any categorization of the sensory signals (top panel of Fig 11.1), predicts Gaussian-like distributions favoring control variables that produce sensory correlates that lie close to the center of the target regions.

A second criterion that would enable to further distinguish between the two variants involving categorical perception modules is best illustrated in the context of adaptation to sensory perturbations (Chapter 8). In this context, auditory and somatosensory predictions do not agree anymore and the values of the control variable planned according to each sensory pathway are shifted from one each other (see Fig 11.1 for illustration). In such case, the variant that combines sensory information before categorization (bottom right panel in Fig 11.1), predicts a plateau distribution corresponding to categorization based on the fusion of sensory pathways (green curve in bottom right panel of Fig 11.1), in which case the width of the plateau does not depend on the position of the distributions predicted by each sensory pathway, but only on their width. On the other hand, the variant that categorizes sensory information before combining them (bottom left panel in Fig 11.1), predicts a plateau distribution that corresponds to the intersection of the categorization plateaus predicted by each sensory pathway [8] (green shaded area in bottom left panel of Fig 11.1). In this case, the width of the resulting plateau critically depends on the position of the distributions predicted by each sensory pathway: the greater the mismatch between sensory pathways, the smaller the resulting intersection of their categorization processes, and hence the smaller the width of the resulting motor planning plateau.

2.2 Caveats and further developments

2.2.1 Integrated model

In this thesis we have developed a series of Bayesian models for speech motor planning. Starting from a Bayesian model for the production of single phonemes, we have extended it in three different directions: towards the inclusion of force constraints, towards sequence planning and towards the description of multisensory targets. Thanks to the modularity of the framework, it is straightforward to combine all three components into a single integrated model. The dependency structure of this integrated model is shown in Fig 11.2. Even though defining this integrated model is conceptually straightforward, it remains to be implemented, simulated, and studied experimentally, in particular concerning the potential interactions between components. For instance, it is likely that differences in patterns of coarticulation may arise with the involvement of somatosensory targets in the planning of phoneme sequences.

2.2.2 Including a mental syllabary and relation with the COSMO model

One of the crucial assumptions of our modeling work is that the motor and phoneme variables, M and Φ, are only related via the sensory or force pathways. In particular, in the multisensory model presented in Chapter 7, motor planning can only be performed by activating at least one of the sensory pathways, otherwise we obtained $P(M \mid \Phi) = P(M)$ (neglecting the pathway related to force), which was not informative at all.

[8] It is important to have in mind that the red and blue distributions in Fig 11.1 correspond to the same category but inferred from two different sensory spaces.

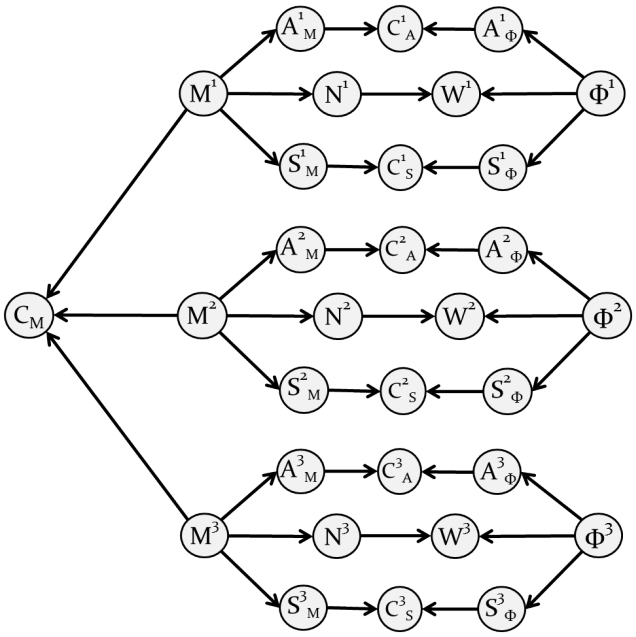

Figure 11.2: Graphical representation of the dependency structure of the integrated model including all components explored in this thesis.

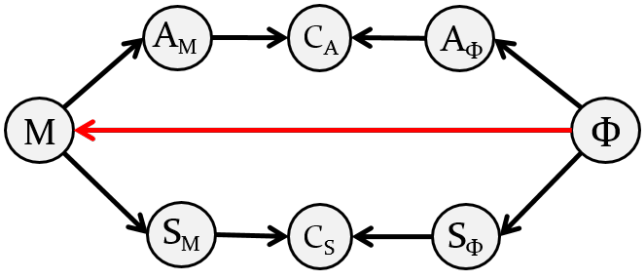

Figure 11.3: Diagram of the multisensory model including an additional "mental syllabary" component (red arrow). *The effort constraint is not indicated for simplicity.*

A possible extension of the model would be to consider that the speech motor system learns a direct relation between patterns of control variables and phonemes, such that $P(M \mid \Phi)$ contains the information obtained from motor planning inferred from the sensory pathways ($P(M \mid \Phi\ [C_A = 1][C_S = 1]$) in the model of Chapter 7, neglecting the effort constraint). This additional knowledge in the model would implement the idea that, for frequently used speech units or sequences of speech units, sets of motor variables adapted to the production of these units could be stored and directly accessed during speech production, from the specification of the units to be produced, without going through the planning process based on sensory pathways. A well-known example of such a storage of motor information related to speech units is the concept of "mental syllabary", proposed by Levelt and Wheeldon (1994), in which motor programs underlying the production of frequently used syllables would be stored. The addition of a "mental-syllabary-like" structure in the model is illustrated by the red arrow linking phoneme and motor variables in Fig 11.3.

Coherence variables also provide a way to include stored motor information related to phonemes in the model, by adding a phoneme variable Φ_M related to the control variable M, and a coherence variable C imposing the coherence between duplicated phoneme variables

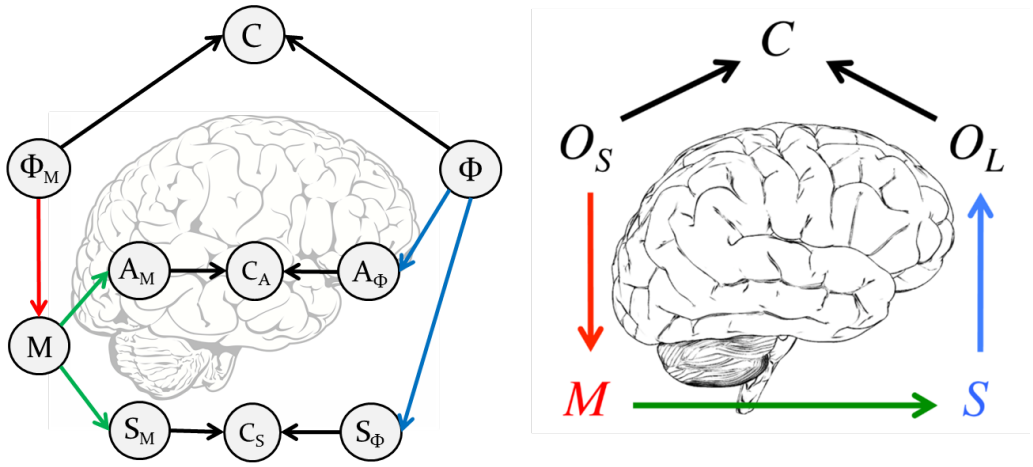

Figure 11.4: The Bayesian model presented in this thesis (left panel), augmented with a mental syllabary, and the COSMO model (right panel) from Barnaud (2018). *Variables* Φ_M *and* Φ_M *on the left panel correspond to variables* O_S *and* O_L *on the right panel. Arrow colors in both panels match and highlight corresponding pieces of models.*

Φ and Φ_M, which can be interpreted as respectively characterizing the phoneme to be produced and the phoneme to be perceived. As suggested in Fig 11.4, this alternative architecture further enables to explicitly relate the modeling work presented in this thesis with the COSMO model family developed in our lab (Barnaud, 2018; Laurent, Barnaud, Schwartz, Bessière, & Diard, 2017; Moulin-Frier, Diard, Schwartz, & Bessière, 2015; Moulin-Frier, Laurent, Bessière, Schwartz, & Diard, 2012). Speech units in the COSMO model are represented by variables O_S and O_L, related to the objects of communication as seen from a speaker or a listener point of view. In the present case we consider phonemes as objects of communication, hence variables O_S and O_L correspond to variables Φ_M and Φ. As can be seen in Fig 11.4, a major difference between our model and the COSMO model is that COSMO considers a single sensory variable S, which we have further specified in terms of auditory and somatosensory variables in the multisensory model presented in Chapter 7. Another important difference arises from the fact that COSMO has been developed with respect to a geometrical model of the vocal tract (the VLAM model (Maeda, 1990)): the motor variables in COSMO are then identified with articulatory parameters, i.e. somehow articulatory positions, that we would rather interpret as somatosensory information in our model. This thesis could thus contribute to the development of the sensory-motor pathway in the COSMO model by (1) specifying the sensory branch in COSMO in terms of somatosensory and auditory pathways, and (2) specifying the motor variable M in COSMO in terms of a more realistic biomechanical description of the speech apparatus.

Furthermore, casting our family of models into the COSMO framework would enable considering some of the theoretical results of COSMO in the context of our model of speech motor planning. For instance, we have hinted several times at a possible mechanism by which one portion of the model would learn its distribution as the result of inference through other portions of the model: we have assumed that somatosensory characterizations of phonemes would summarize the result of acoustic-based planning; we have suggested that a mental-syllabary-like structure would summarize the result of sensory-based planning. In the COSMO framework, indistinguishability theorems describe asymptotic results where one portion of the model learns another portion so well that their information content become identical, and they become experimentally indistinguishable. This was studied, in COSMO, in the context of auditory and motor models of speech perception; in our case, this would open the way towards equivalent

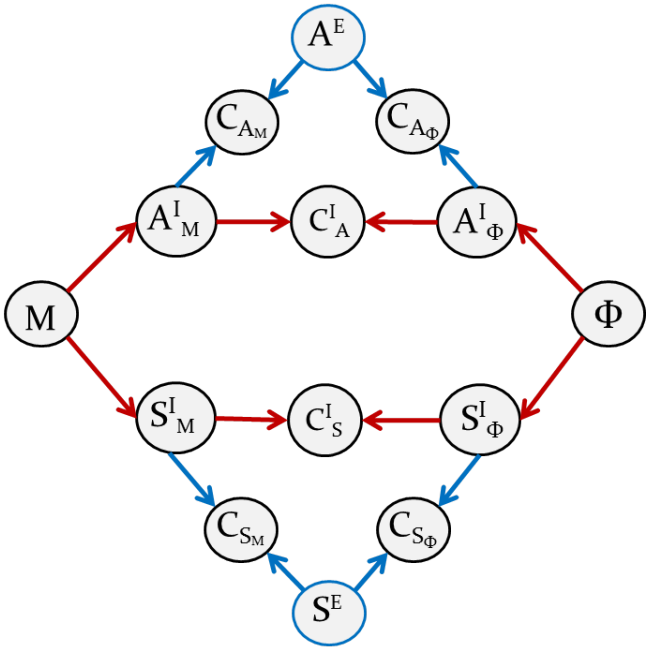

Figure 11.5: Including external (or actual, as opposed to internal) sensory feedback in the model. *Red arrows indicate dependencies between variables as proposed in the model presented in Chapter 7. Blue arrows indicate dependencies between additional sensory variables corresponding to external sensory feedback (variables A^E and S^E).*

"indistinguishability theorems" in the context of speech production. Our study of speech motor adaptation (Chapters 8 and 10) further suggests that such indistinguishability relations may be temporarily disrupted in perturbation experiments; this would suggest that adaptation might be interpreted as a mechanism for recovering indistinguishability. Such an intriguing possibility would open new perspectives for the study of adaptation mechanisms and dynamics.

2.2.3 Including sensory feedback and time

Finally, a major limitation of the present modeling work is that it does not take time into account and does not include actual sensory feedback in the planning, but only internal sensory predictions. Indeed, we have worked on a model for motor planning that is part of a control strategy that separates planning from execution. Currently, time is only taken into account during execution, downstream, during the control of the biomechanical model, and sensory feedback is only integrated at a very low level, namely via an equivalent of the muscle stretch reflex. There are two main reasons for which it would be essential to further develop the present work in order to account for time and integrate sensory feedback (in addition to internal sensory predictions) at the planning level. Firstly, while the planning-execution dichotomy may be well suited to describe the fast and skilled patterns of speech in normal conditions (such as would be planned, in an automatic manner in nominal conditions by a mental syllabary), it is seriously challenged by the online (yet delayed) compensatory behavior observed in the context of unexpected auditory perturbations (Burnett & Larson, 2002; Purcell & Munhall, 2006; Tourville, Reilly, & Guenther, 2008). It would be crucial to further develop the model in order to account for correction and re-planning mechanisms during execution that would enable to modify the temporal pattern of control variables in the presence of perturbations, or recompute their values altogether.

The second important reason that requires accounting for time and sensory feedback at the planning level is the study of the dynamic of adaptation to sensory perturbations during the training phase. Although we have explored hypotheses of possible changes resulting from adaptation to sensory perturbations (Chapters 8 and 10), we did not investigate how sensory errors are integrated during the process of adaptation and what learning mechanisms lead to the update of the different pieces of knowledge included in the model. In particular it would be interesting to explore how different magnitudes of sensory mismatch may lead to different dynamics of adaptation, in line with similar studies developed in the context of arm reaching or bird song (Hahnloser & Narula, 2017; Sober & Brainard, 2012; Wei & Kording, 2009).

The first step in a development of such a model would be to include variables corresponding to external sensory feedback in addition to variables corresponding to internal sensory predictions. Again, coherence variables would be of great help in this context. A possible model architecture including sensory feedback with coherence variables is illustrated in Fig 11.5. In this model, variables corresponding to external sensory feedback are denoted by A^E and S^E, and are connected to internal sensory variables (corresponding to sensory-motor predictions, A^I_M and S^I_M, and to sensory expectations characterizing phonemes, A^I_Φ and S^I_Φ), with four additional coherence variables that enable to connect or disconnect sensory feedback from internal predictions.

Note that currently the use of coherence variables in our models has been limited to mathematical switches, activating or not the coherence constraint connecting sensory pathways in the model. That is, in inference, we have assumed values of coherence variables, thus setting the states of these connector variables. However, coherence variables can also be targets of inference, implementing an error detection mechanism (Phénix, 2018). Inference, in this case, evaluates the probability that a given coherence variable is "closed", that is to say, that variables on either side match, relative to the matching constraint that this coherence variable provides (which we have shown to be fully mathematically controllable, Chapter 8). Such an error detection mechanism could be further applied to the context of sensorimotor learning, generating learning signals, provoking updates of the sensory-motor internal models, $P(A_M \mid M)$ and $P(S_M \mid M)$, and of the sensory characterization of phonemes, $P(A_\Phi \mid \Phi)$ and $P(S_\Phi \mid \Phi)$. Further developing the model in this direction and relating such ideas to classical dynamical system architectures, where hierarchical feedback loops and response delays are commonly used (Houde & Nagarajan, 2011), or to other recent developments in the field of predictive coding and active inference (Adams, Shipp, & Friston, 2013; Friston, 2010; Friston et al., 2016), would be two very exciting endeavors.

Bibliography

Adams, R. A., Shipp, S., & Friston, K. J. (2013). Predictions not commands: active inference in the motor system. *Brain Structure and Function*, *218*(3), 611–643.

Badin, P., & Fant, G. (1984). Notes on vocal tract computation. *STL QPSR*, *2*(3), 53–108.

Baer, T., Alfonso, P., & Honda, K. (1988). Electromyography of the tongue muscles during vowels in /ɔpʊp/ environment. *Ann. Bull. RILP*, *22*, 7–19.

Balasubramaniam, R., & Feldman, A. G. (2004). Guiding movements without redundancy problems. In *Coordination dynamics: Issues and trends* (pp. 155–176). Springer.

Barnaud, M.-L. (2018). *Modélisation bayésienne du développement conjoint de la perception, l'action et la phonologie* (Unpublished doctoral dissertation). Univ. Grenoble Alpes.

Bazzoli, C., Letué, F., & Martinez, M.-J. (2015). Modelling finger force produced from different tasks using linear mixed models with lme R function. *Journal of Case Studies in Business, Industry and Government Statistics (CSBIGS)*, *6*(1), 16–36.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron*, *74*(1), 30–39.

Beckman, M. E., Jung, T.-P., Lee, S.-H., de Jong, K., Krishnamurthy, A. K., Ahalt, S. C., ... Collins, M. J. (1995). Variability in the production of quantal vowels revisited. *The Journal of the Acoustical Society of America*, *97*(1), 471–490.

Bernstein, N. A. (1967). *The control and regulation of movements.* London: Pergamon Press.

Bessière, P., Laugier, C., & Siegwart, R. (Eds.). (2008). *Probabilistic reasoning and decision making in sensory-motor systems* (Vol. 46). Berlin: Springer-Verlag.

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

Bishop, M., C. (2006). *Pattern recognition and machine learning.* New York, NY: Springer.

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, *66*(4), 1001–1017.

Brand, R. A., Pedersen, D. R., & Friederich, J. A. (1986). The sensitivity of muscle force predictions to changes in physiologic cross-sectional area. *Journal of Biomechanics*, *19*(8), 589–596.

Browman, C. P., & Goldstein, L. (1985). Dynamic modeling of phonetic structure. In F. V. A. (Ed.), *Phonetic linguistics* (pp. 35–53). New York: Academic Press.

Browman, C. P., & Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica*, *45*(2-4), 140–155.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3-4), 155–180.

Brown, L. D. (1981). A complete class theorem for statistical problems with finite sample spaces. *The Annals of Statistics*, 1289–1300.

Brunner, J., Hoole, P., & Perrier, P. (2011). Adaptation strategies in perturbed/s/. *Clinical linguistics & phonetics*, *25*(8), 705–724.

Buchaillard, S., Perrier, P., & Payan, Y. (2009). A biomechanical model of cardinal vowel production: muscle activations and the impact of gravity on tongue positioning. *The Journal of the Acoustical Society of America*, *126*(4), 2033–2051.

Burnett, T. A., Freedland, M. B., & Larson, C. R. (1998). Voice F0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America*, *103*(6), 3153–3161.

Burnett, T. A., & Larson, C. R. (2002). Early pitch-shift response is active in both steady and dynamic voice pitch control. *The Journal of the Acoustical Society of America*, *112*(3), 1058–1063.

Cai, S., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2010). Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong /iau/ and its pattern of generalization. *The Journal of the Acoustical Society of America*, *128*(4), 2033–48. doi: 10.1121/1.3479539

Calliope. (1984). *La parole et son traitement automatique*. Masson.

Castellanos, A., Benedí, J.-M., & Casacuberta, F. (1996). An analysis of general acoustic-phonetic features for spanish speech produced with the lombard effect. *Speech Communication*, *20*(1-2), 23–35.

Churchland, M. M., Afshar, A., & Shenoy, K. V. (2006). A central source of movement variability. *Neuron*, *52*(6), 1085–1096.

Colas, F., Diard, J., & Bessière, P. (2010). Common Bayesian models for common cognitive issues. *Acta Biotheoretica*, *58*(2-3), 191–216.

Combescure, P. (1981). 20 listes de dix phrases phonétiquement équilibrées. *Revue d'Acoustique*, *56*, 34–38.

Crevier-Buchman, L., Gendrot, C., Denby, B., Pillot-Loiseau, C., Roussel, P., Colazo-Simon, A., & Dreyfus, G. (2011). Articulatory strategies for lip and tongue movements in silent versus vocalized speech. In *17th International Congress of Phonetic Science (ICPhS) 2011* (pp. 1–4).

Dang, J., & Honda, K. (2004). Construction and control of a physiological articulatory model. *The Journal of the Acoustical Society of America*, *115*(2), 853–870.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, *93*(3), 283.

Diard, J. (2015). *Bayesian Algorithmic Modeling in Cognitive Science* (Habilitation à Diriger des Recherches (HDR), Université Grenoble Alpes). doi: 10.13140/RG.2.1.1756.0404

Dromey, C., & Black, K. M. (2017). Effects of laryngeal activity on articulation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(12), 2272–2280.

Egan, J. J. (1972). Psychoacoustics of the lombard voice response. *Journal of Auditory Research*.

Epps, J., Smith, J., & Wolfe, J. (1997). A novel instrument to measure acoustic resonances of the vocal tract during phonation. *Measurement Science and Technology*, *8*(10), 1112.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. doi: 10.1038/415429a

Fabre, D. (2016). *Retour articulatoire visuel par échographie linguale augmentée: développements et application clinique* (Unpublished doctoral dissertation). Université Grenoble Alpes.

Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature reviews neuroscience*, *9*(4), 292.

Feldman, A. G. (1986). Once more on the equilibrium-point hypothesis (λ model) for motor control. *Journal of motor behavior*, *18*(1), 17–54.

Feng, Y., Gracco, V. L., & Max, L. (2011). Integration of auditory and somatosensory error signals in the neural control of speech movements. *Journal of neurophysiology*, *106*(2), 667–679.

Folkins, J. W., & Brown, C. K. (1987). Upper lip, lower lip, and jaw interactions during

speech: Comments on evidence from repetition-to-repetition variability. *The Journal of the Acoustical Society of America*, *82*(6), 1919–1924.

Fowler, C. A. (1991). Auditory perception is not special: We see the world, we feel the world, we hear the world. *The Journal of the Acoustical Society of America*, *89*(6), 2910–2915.

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *The Journal of the Acoustical Society of America*, *99*(3), 1730–1741. doi: 10.1121/1.415237

Fowler, C. A., & Turvey, M. T. (1980). Immediate compensation in bite-block speech. *Phonetica*, *37*(5-6), 306–326.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., Pezzulo, G., et al. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, *68*, 862–879.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic bulletin & review*, *13*(3), 361–377.

Ganesh, G., Haruno, M., Kawato, M., & Burdet, E. (2010). Motor memory and local minimization of error and effort, not global optimization, determine motor behavior. *Journal of neurophysiology*, *104*(1), 382–390.

Gay, T., Lindblom, B., & Lubker, J. (1981). Production of bite-block vowels: Acoustic equivalence by selective compensation. *The Journal of the Acoustical Society of America*, *69*(3), 802–810.

Gérard, J.-M., Wilhelms-Tricarico, R., Perrier, P., & Payan, Y. (2003). A 3d dynamical biomechanical tongue model to study speech motor control. *Recent Research Developments in Biomechanics*, 49–64.

Ghez, C., Scheidt, R., & Heijink, H. (2007). Different learned coordinate frames for planning trajectories and final positions in reaching. *Journal of neurophysiology*, *98*(6), 3614–3626.

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.

Gordon, J., Ghilardi, M. F., Cooper, S. E., & Ghez, C. (1994). Accuracy of planar reaching movements. *Experimental brain research*, *99*(1), 112–130.

Grimme, B., Fuchs, S., Perrier, P., & Schöner, G. (2011). Limb versus speech motor control: A conceptual review. *Motor control*, *15*(1), 5–33.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological review*, *102*(3), 594–621.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language*, *96*(3), 280–301.

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological review*, *105*(4), 611.

Hahnloser, R. H., & Narula, G. (2017). A bayesian account of vocal adaptation to pitch-shifted auditory feedback. *PloS one*, *12*(1), e0169795.

Haith, A., Jackson, C. P., Miall, R. C., & Vijayakumar, S. (2009). Unifying the sensory and motor components of sensorimotor adaptation. In *Advances in Neural Information Processing Systems* (pp. 593–600).

Haith, A., & Krakauer, J. (2013). Theoretical models of motor control and motor learning. *Routledge handbook of motor control and motor learning*, 1–28.

Halle, M., & Chomsky, N. (1968). *The sound pattern of english.* Harper & Row.

Hannam, A. G., Stavness, I., Lloyd, J. E., & Fels, S. (2008). A dynamic model of jaw and hyoid biomechanics during chewing. *Journal of Biomechanics*, *41*(5), 1069–1076.

Harris, C. M., & Wolpert, D. M. (1998). Signal-dependent noise determines motor planning.

Nature, *394*(6695), 780–784.

Harshman, R., Ladefoged, P., & Goldstein, L. (1977). Factor analysis of tongue shapes. *The Journal of the Acoustical Society of America*, *62*(3), 693–707.

Honda, K. (1996). Organization of tongue articulation for vowels. *Journal of Phonetics*, *24*, 39–52.

Honda, M., Fujino, A., & Kaburagi, T. (2002). Compensatory responses of articulators to unexpected perturbation of the palate shape. *Journal of phonetics*, *30*(3), 281–302.

Hoole, P. (1998). Modelling tongue configuration in german vowel production. In *Fifth International Conference on Spoken Language Processing.*

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical journal*, *50*(3), 346–363.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor Adaptation in Speech Production. *Science (New York, N.Y.)*, *1*(5354), 1213–1216. doi: 10.1126/science.279.5354.1213

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in human neuroscience*, *5*.

Hueber, T., Badin, P., Savariaux, C., Vilain, C., & Bailly, G. (2010). Differences in articulatory strategies between silent, whispered and normal speech ? a pilot study using electromagnetic articulography. In *Proceedings of International Seminar on Speech Production.* Montréal Canada.

Jackson, M. T. (1988). Analysis of tongue positions: Language-specific and cross-linguistic models. *The Journal of the Acoustical Society of America*, *84*(1), 124–143.

Janke, M., Wand, M., & Schultz, T. (2010). Impact of lack of acoustic feedback in emg-based silent speech recognition. In *Eleventh Annual Conference of the International Speech Communication Association.*

Jaynes, E. T. (2003). *Probability theory: The logic of science* (G. L. Bretthorst, Ed.). Cambridge, UK: Cambridge University Press.

Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of F0 production: evidence from feedback perturbation. *The Journal of the Acoustical Society of America*, *108*(3), 1246–1251.

Jordan, I., M., & Rumelhart, E., D. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, *16*(3), 307—354.

Junqua, J.-C. (1993). The lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, *93*(1), 510–524.

Kakita, Y., Fujimura, O., & Honda, K. (1985). Computation of Mapping from Muscular Contraction Patterns to Formant Patterns in Vowel Space. In V. Fromkin (Ed.), *Phonetic Linguistics: Essays in Honour of Peter Ladefoged* (Vol. 74, p. 133-144). Orlando: Academic Press.

Kallail, K., & Emanuel, F. (1985). The identifiability of isolated whispered and phonated vowel samples. *Journal of phonetics*, *13*(1), 11–17.

Katseff, S., Houde, J., & Johnson, K. (2012). Partial compensation for altered auditory feedback: A tradeoff with somatosensory feedback? *Language and Speech*, *55*(2), 295–308. doi: 10.1177/0023830911417802

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current opinion in neurobiology*, *9*(6), 718–727.

Kelly, J. L. (1973). Speech synthesis. In *Proc. 4th Int. Congr. Acoustics* (pp. 1–4).

Kiritani, S. (1976). A computational model of the tongue. *Ann Bull RILP*, *10*, 243–251.

Lametti, D. R., Nasir, S. M., & Ostry, D. J. (2012). Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. *The Journal*

of neuroscience : the official journal of the Society for Neuroscience, *32*(27), 9351–9358. doi: 10.1523/JNEUROSCI.0404-12.2012

Lametti, D. R., & Ostry, D. J. (2010). Postural constraints on movement variability. *Journal of Neurophysiology*, *104*(2), 1061–1067.

Lametti, D. R., Rochet-Capellan, A., Neufeld, E., Shiller, D. M., & Ostry, D. J. (2014). Plasticity in the Human Speech Motor System Drives Changes in Speech Perception. *Journal of Neuroscience*, *34*(31), 10339–10346. doi: 10.1523/JNEUROSCI.0108-14.2014

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Research*, *35*(3), 389–412.

Lane, H., & Tranel, B. (1971). The lombard sign and the role of hearing in speech. *Journal of Speech, Language, and Hearing Research*, *14*(4), 677–709.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (p. 112-136). New York: Wiley.

Latash, M. (2010). Motor control: in search of physics of the living systems. *Journal of Human Kinetics*, *24*, 7–18.

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*.

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

Lechner, B. K. (1979). The effects of delayed auditory feedback and masking on the fundamental frequency of stutterers and nonstutterers. *Journal of Speech, Language, and Hearing Research*, *22*(2), 343–353.

Levelt, W. J., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition*, *50*(1-3), 239–269.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36. doi: 10.1016/0010-0277(85)90021-6

Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science*, *243*(4890), 489–494.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the h&h theory. In *Speech Production and Speech Modelling* (pp. 403–439). Springer.

Lombard, E. (1911). Le signe de l'élévation de la voix. *Annales des Maladies de L'Oreille et du Larynx*, *37*, 101–119.

Lubker, J. F. (1979). The reorganization times of bite-block vowels. *Phonetica*, *36*(4-5), 273–293.

Ma, L., Perrier, P., & Dang, J. (2015). Strength of syllabic influences on articulation in mandarin chinese and french: Insights from a motor control approach. *Journal of Phonetics*, *53*, 101–124.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech Production and Speech Modelling* (pp. 131–149). Springer.

Marr, D. (1982). *Vision. a computational investigation into the human representation and processing of visual information.* New York, USA: W.H. Freeman and Company.

Ménard, L. (2002). *Production et perception des voyelles au cours de la croissance du conduit vocal : variabilité, invariance et normalisation* (Unpublished Ph.D. Thesis). Université Stendhal de Grenoble.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America*, *53*(4), 1070–1082.

Mooshammer, C., Perrier, P., Fuchs, S., Geng, C., & Pape, D. (2004). An EMMA and EPG study on token-to-token variability. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel*, *36*(46–63).

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.

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. doi: 10.1080/01690965.2011.645313

Mundry, R., & Nunn, C. L. (2009). Stepwise model fitting and statistical inference: turning noise into signal pollution. *The American Naturalist*, *173*(1), 119–123.

Nam, H., Goldstein, L., & Saltzman, E. L. (2009). Self-organization of syllable structure: A coupled oscillator model. In F. Pellegrino, M. Marsico, I. Chitoran, & C. Coupé (Eds.), *Approaches to phonological complexity, Phonology and Phonetic Series 16* (pp. 299–328). Mouton de Gruyter Berlin, Germany, New York, NY.

Nasir, S. M., & Ostry, D. J. (2009). Auditory plasticity and speech motor learning. *Proceedings of the National Academy of Sciences*, *106*(48), 20470–20475.

Nazari, M. A., Perrier, P., Chabanas, M., & Payan, Y. (2010). Simulation of dynamic orofacial movements using a constitutive law varying with muscle activation. *Computer Methods in Biomechanics and Biomedical Engineering*, *13*(4), 469–482.

Nelson, W. (1983). Physical principles for economies of skilled movements. *Biological Cybernetics*, *46*, 135—147.

Nix, D. A., Papcun, G., Hogden, J., & Zlokarnik, I. (1996). Two cross-linguistic factors underlying tongue shapes for vowels. *The Journal of the Acoustical Society of America*, *99*(6), 3707–3717.

Osborne, L. C., Hohl, S. S., Bialek, W., & Lisberger, S. G. (2007). Time course of precision in smooth-pursuit eye movements of monkeys. *Journal of Neuroscience*, *27*(11), 2987–2998.

Osborne, L. C., Lisberger, S. G., & Bialek, W. (2005). A sensory source for motor variation. *Nature*, *437*(7057), 412.

Ostry, D. J., Darainy, M., Mattar, A. A. G., Wong, J., & Gribble, P. L. (2010). Somatosensory plasticity and motor learning. *The Journal of Neuroscience*, *30*(15), 5384–93.

Ouni, S. (2014). Tongue control and its implication in pronunciation training. *Computer Assisted Language Learning*, *27*(5), 439–453.

Parrell, B., Ramanarayanan, V., Nagarajan, S., & Houde, J. (2018). FACTS: A hierarchical task-based control model of speech incorporatingsensory feedback. In *Interspeech 2018* (p. Submitted).

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.

Patri, J.-F., Diard, J., & Perrier, P. (2016). Modélisation bayésienne de la planification motrice des gestes de parole : Évaluation du rôle des différentes modalités sensorielles. In *J.E.P. 2016* (Vol. 1, pp. 419–427).

Patri, J.-F., Diard, J., & Perrier, P. (2017). Modeling sensory preference in speech motor planning. In *Neural Control of Movement.* Dublin, Ireland.

Patri, J.-F., Diard, J., Schwartz, J.-L., & Perrier, P. (2015a). A Bayesian framework for speech motor control. In *Progress in Motor Control X.* Budapest, Hungary.

Patri, J.-F., Diard, J., Schwartz, J.-L., & Perrier, P. (2015b). A Bayesian framework for speech

motor control. In *Workshop "Probabilistic Inference and the Brain"*. Paris, France.

Patri, J.-F., Perrier, P., & Diard, J. (2016). Bayesian modeling in speech motor control: a principled structure for the integration of various constraints. In *Interspeech 2016* (pp. 3588–3592). San Francisco. doi: 10.21437/Interspeech.2016-441

Patri, J.-F., Perrier, P., & Diard, J. (2017). Modeling sensory preference in speech motor planning. In *Proceedings of the 11th International Seminar on Speech Production (ISSP 2017)*. Tianjin, China.

Patri, J.-F., Perrier, P., Schwartz, J.-L., & Diard, J. (2018). What drives the perceptual change resulting from speech motor adaptation? evaluation of hypotheses in a bayesian modeling framework. *PLoS Computational Biology*, *14*(1), e1005942.

Payan, Y., & Perrier, P. (1997). Synthesis of VV sequences with a 2D biomechanical tongue model controlled by the equilibrium point hypothesis. *Speech communication*, *22*(2), 185–205.

Pelteret, J. P., & Reddy, B. D. (2012). Computational model of soft tissues in the human upper airway. *International journal for numerical methods in biomedical engineering*, *28*(1), 111–132.

Perkell, J. S. (1974). *A physiologically-oriented model of tongue activity in speech production.* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Perkell, J. S. (1980). Phonetic features and the physiology of speech production. *Language Production Vol. 1: Speech and Talk*, 337–372.

Perkell, J. S. (2013). Five decades of research in speech motor control: What have we learned, and where should we go from here? *Journal of Speech, Language, and Hearing Research*, *56*(6), S1857–S1874.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Perrier, P., Vick, J., ... Zandipour, M. (2000). A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonetics*, *28*(3), 233–272.

Perkell, J. S., & Klatt, D. H. (2014). *Invariance and variability in speech processes.* Psychology Press.

Perkell, J. S., Lane, H., Ghosh, S., & Matthies, M. L. (2008). Mechanisms of Vowel Production: Auditory Goals and Speaker Acuity. In *Proceedings of the 8th International Seminar on Speech Production, Strasbourg, France* (pp. 29–32).

Perkell, S., J., & Nelson, L., W. (1985). Variability in production of the vowels /i/ and /a/. *Journal of the Acoustical Society of America*, *77*, 1889–1895.

Perrier, P., Boë, L.-J., & Sock, R. (1992). Vocal tract area function estimation from midsagittal dimensions with ct scans and a vocal tract cast: modeling the transition with two sets of coefficients. *Journal of Speech, Language, and Hearing Research*, *35*(1), 53–67.

Perrier, P., & Fuchs, S. (2008). Speed–curvature relations in speech production challenge the 1/3 power law. *Journal of Neurophysiology*, *100*(3), 1171–1183.

Perrier, P., & Fuchs, S. (2015). Motor equivalence in speech production. In *The handbook of speech production* (pp. 225–247). John Wiley & Sons.

Perrier, P., Lœvenbruck, H., & Payan, Y. (1996). Control of tongue movements in speech: The equilibrium point hypothesis perspective. *Journal of Phonetics*, *24*(1), 53–75.

Perrier, P., & Ma, L. (2008). Speech planning for V1CV2 sequences: Influence of the planned sequence. In ISSP-2008 (Ed.), *Proceedings of the 8th International Seminar on Speech Production (ISSP 2008)* (pp. 69–72). Université de Strasbourg, France.

Perrier, P., Ma, L., & Payan, Y. (2005). Modeling the production of VCV sequences via the inversion of a biomechanical model of the tongue. In *Proceedings of Interspeech 2005* (p. 1041-–1044). Lisbon, Portugal.

Perrier, P., Payan, Y., Zandipour, M., & Perkell, J. (2003). Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *The Journal of the Acoustical Society of America*, *114*(3), 1582–1599.

Phénix, T. (2018). *Modélisation bayésienne algorithmique de la reconnaissance visuelle de mots et de l'attention visuelle* (Unpublished doctoral dissertation). Univ. Grenoble Alpes.

Pijpers, M., Alder, M. D., & Togneri, R. (1993). Finding structure in the vowel space. In *First Australian and New Zealand Conference on Intelligent Information Systems.* Menlo Park, CA: IEEE Press.

Poggio, T., & Girosi, F. (1989). *A theory of networks for approximation and learning* (Tech. Rep.). Cambridge, MA, USA: Artificial Intelligence Laboratory & Center for Biological Information Processing, MIT.

Pruim, G., De Jongh, H., & Ten Bosch, J. (1980). Forces acting on the mandible during bilateral static bite at different bite force levels. *Journal of biomechanics*, *13*(9), 755–763.

Purcell, D. W., & Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *The Journal of the Acoustical Society of America*, *119*(4), 2288–2297. doi: 10.1121/1.2173514

Ramanarayanan, V., Parrell, B., Goldstein, L., Nagarajan, S., & Houde, J. (2016). A new model of speech motor control based on task dynamics and state feedback. In *Interspeech* (pp. 3564–3568).

Robert, C. P. (2007). *The Bayesian choice – from decision-theoretic foundations to computational implementation.* Springer.

Robert-Ribes, J. (1995). *Modèles d'intégration audiovisuelle de signaux linguistiques : de la perception humaine a la reconnaissance automatique des voyelles* (Unpublished Ph.D. Thesis). Institut National Polytechnique de Grenoble.

Rosenbaum, D. A., Loukopoulos, L. D., Meulenbroek, R. G., Vaughan, J., & Engelbrecht, S. E. (1995). Planning reaches by evaluating stored postures. *Psychological review*, *102*(1), 28.

Rubin, P., Saltzman, E. L., Goldstein, L., McGowan, R., Tiede, M., & Browman, C. (1996). Casy and extensions to the task-dynamic model. In *1st ETRW on Speech Production Modeling: From Control Strategies to Acoustics; 4th Speech Production Seminar: Models and Data.*

Saltzman, E. L. (1986). Task dynamic coordination of the speech articulators: A preliminary model. In H. Heuer & C. Fromm (Eds.), *Generation and Modulation of Action Patterns* (Experimental Brain Research Series ed., pp. 129–144). New York: Springer-Verlag.

Saltzman, E. L. (1991). The task dynamic model in speech production. *Speech motor control and stuttering*, 37–52.

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological psychology*, *1*(4), 333–382.

Savariaux, C., Badin, P., Samson, A., & Gerber, S. (2017). A comparative study of the precision of carstens and northern digital instruments electromagnetic articulographs. *Journal of Speech, Language, and Hearing Research*, *60*(2), 322–340.

Savariaux, C., Perrier, P., & Orliaguet, J.-P. (1995, nov). Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. *The Journal of the Acoustical Society of America*, *98*(5), 2428–2442. doi: 10.1121/1.413277

Schuerman, W. L., Nagarajan, S., McQueen, J. M., & Houde, J. (2017). Sensorimotor adaptation affects perceptual compensation for coarticulation. *The Journal of the Acoustical Society of America*, *141*(4), 2693–2704.

Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. (2012). The perception-for-action-control theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, *25*(5), 336–354.

Schwartz, J.-L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after summerfield: a taxonomy of models for audio-visual fusion in speech perception. *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech*, 85–108.

Shadmehr, R., & Mussa-Ivaldi, S. (2012). *Biological learning and control: how the brain builds representations, predicts events, and makes decisions*. Mit Press.

Shiller, D. M., Sato, M., Gracco, V. L., & Baum, S. R. (2009). Perceptual recalibration of speech sounds following speech motor learning. *The Journal of the Acoustical Society of America*, *125*(2), 1103–1113. doi: 10.1121/1.3058638

Shim, J. K., Latash, M. L., & Zatsiorsky, V. M. (2003). Prehension synergies: trial-to-trial variability and hierarchical organization of stable performance. *Experimental Brain Research*, *152*(2), 173–184.

Sober, S. J., & Brainard, M. S. (2012). Vocal learning is constrained by the statistics of sensorimotor experience. *Proceedings of the National Academy of Sciences*, *109*(51), 21099–21103.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In e. David Jr. E.E. & Denes P.B. (Ed.), *Human Communication: A unified view* (pp. 51–66). New York: McGraw-Hill.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, *17*, 3–45. doi: 10.1109/ICSLP.1996.607202

Story, B. H., Laukkanen, A.-M., & Titze, I. R. (2000). Acoustic impedance of an artificially lengthened and constricted vocal tract. *Journal of Voice*, *14*(4), 455–469.

Story, B. H., & Titze, I. R. (1995). Voice simulation with a body-cover model of the vocal folds. *The Journal of the Acoustical Society of America*, *97*(2), 1249–1260.

Stuart, A., Kalinowski, J., Rastatter, M. P., & Lynch, K. (2002). Effect of delayed auditory feedback on normal speakers at two speech rates. *The Journal of the Acoustical Society of America*, *111*(5), 2237–2241.

Szabados, A. (2017). *Uncontrolled manifolds et réflexes à courte latence dans le contrôle moteur de la parole: une étude de modélisation* (Unpublished doctoral dissertation). Univ. Grenoble Alpes.

Tartter, V. C. (1991). Identifiability of vowels and speakers from whispered syllables. *Perception & psychophysics*, *49*(4), 365–372.

Tasko, S. M., & Westbury, J. R. (2004). Speed–curvature relations for speech-related articulatory movement. *Journal of Phonetics*, *32*(1), 65–80.

Tilsen, S. (2017). Exertive modulation of speech and articulatory phasing. *Journal of Phonetics*, *64*, 34–50.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature neuroscience*, *7*(9), 907–915.

Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature neuroscience*, *5*(11), 1226–1235.

Tourville, J. A., Reilly, K. J., & Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *Neuroimage*, *39*(3), 1429–1443.

Tremblay, S., Shiller, D. M., & Ostry, D. J. (2003). Somatosensory basis of speech production. *Nature*, *423*(6942), 866–869. doi: 10.1038/nature01710

Turvey, M. T. (1978). Issues in the theory of action: Degree of freedom, coordinative structures and coalitions. *Attention and performance*, 557–595.

Turvey, M. T., Fitch, H. L., & Tuller, B. (1982). The Bernstein perspective: I. The problems of degrees of freedom and context-conditioned variability. *Human motor behavior: An introduction*, 239–252.

van Beers, R. J. (2007). The sources of variability in saccadic eye movements. *Journal of Neuroscience*, *27*(33), 8757–8770.

van Beers, R. J., Haggard, P., & Wolpert, D. M. (2004). The role of execution noise in movement variability. *Journal of neurophysiology*, *91*(2), 1050–1063.

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America*, *122*(4), 2306–2319. doi: 10.1121/1.2773966

Viviani, P., & Flash, T. (1995). Minimum-jerk, two-thirds power law, and isochrony: converging approaches to movement planning. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 32.

Viviani, P., & Terzuolo, C. (1982). Trajectory determines movement dynamics. *Neuroscience*, *7*(2), 431–437.

Waltl, S., & Hoole, P. (2008). An EMG study of the German vowel system. In R. Sock, S. Fuchs, & Y. Laprie (Eds.), *Proceedings of the 8th International Seminar on Speech Production (ISSP 2008)* (pp. 445–448).

Wei, K., & Kording, K. (2009). Relevance of error: what drives motor adaptation? *Journal of neurophysiology*, *101*(2), 655–664.

Whalen, D., Chen, W.-R., Tiede, M. K., & Nam, H. (2018). Variability of articulator positions and formants across nine english vowels. *Journal of Phonetics*, *68*, 1–14.

Wilhelms-Tricarico, R. (1995). Physiological modeling of speech production: Methods for modeling soft-tissue articulators. *The Journal of the Acoustical Society of America*, *97*(5), 3085–3098.

Winkler, R., Ma, L., & Perrier, P. (2011). A model of optimal speech production planning integrating dynamical constraints to achieve appropriate articulatory timing. In ISSP2011 (Ed.), *Proceedings of the 9th International Seminar on Speech Production* (Vol. Abstracts, p. 235—236). Montréal Canada.

Wu, H., Badin, P., Cheng, Y., & Guerin, B. (1987). Continuous variation of the vocal tract length in a kelly-lochbaum type speech production model. *Proc. Int. Congr. Phonetic Sciences (XIth ICPhS), Tallin, Estonia*, 340–343.

Wyart, V., & Koechlin, E. (2016). Choice variability and suboptimality in uncertain environments. *Current Opinion in Behavioral Sciences*, *11*, 109–115.

Yamamoto, K., & Kawabata, H. (2014). Adaptation to delayed auditory feedback induces the temporal recalibration effect in both speech perception and production. *Experimental Brain Research*, *232*(12), 3707–3718.

Yates, A. (1965). Delayed auditory feedback and shadowing. *The Quarterly Journal of Experimental Psychology*, *17*(2), 125–131. doi: 10.1080/17470216508416421

Appendix A

GEPPETO as a special case of the Bayesian model

The probability distribution $P(M^{1:3} \mid \Phi^{1:3}\ [C_m = L])$ characterizes the set of every sequence of control variables $M^{1:3}$ with its probability to achieve the desired sequence of phoneme $\Phi^{1:3}$ with the "minimum effort" constraint ($C_m = L$). If we look for the most probable solutions, the Bayesian model become equivalent to the optimal control approach as we will now see. This equivalence comes from the fact that the cost function optimized by GEPPETO is proportional to the negative log probability of the inferred control variables. Thus, finding the more probable control parameters under the Bayesian model is equivalent to minimizing the cost function defined by GEPPETO. For simplicity we will derive this proof for a sequence of two phonemes, the result for a sequence of three phonemes being easily obtained in the same way.

For a sequence of two phonemes we have

$$\begin{aligned} & P(M^{1:2} \mid \Phi^{1:2}\ [C_m = L]) \\ & \quad \propto \quad P(\Phi^1 \mid S^*(M^1))\ P(\Phi^2 \mid S^*(M^2))\ e^{-|M^2 - M^1|}, \end{aligned} \tag{A.1}$$

where the parameter κ_M has been set to 1. Let us rewrite Equation (A.1) in the following form

$$P(M^{1:2} \mid \Phi^{1:2}\ [C_m = L]) \quad \propto \quad e^{-\mathcal{L}_B} \tag{A.2}$$

where

$$\mathcal{L}_B \quad = \quad -\ln P(\Phi^1 \mid S^*(M^1)) - \ln P(\Phi^2 \mid S^*(M^2)) + |M^2 - M^1|. \tag{A.3}$$

As will appears shortly $\mathcal{L}_B$ can bee seen as a Lagrangian associated to the Bayesian model. We now unpack the form of this expression in order to compare it to GEPPETO.

It can already be noted that the last term in Equation (A.3) corresponds to the cost function of GEPPETO, as it is just the Euclidean distance between the two motor control variables M^1 and M^2.

Since the form of the distributions $P(\Phi \mid S)$ are close to step functions on the elliptic domains defined for each phoneme (Figure 4 and 5 of the main text), we approximate them as:

$$P(\Phi^i \mid S^*(M^i)) \quad = \quad \begin{cases} 1 & \text{if } S^*(M^i) \in \mathcal{E}_{\Phi^i} \\ 0 & \text{otherwise} \end{cases} \tag{A.4}$$

where $S^*(M^i)$ stands for the spectral characteristics of the acoustic signal corresponding to M^i and $\mathcal{E}^i_\Phi$ correspond to the elliptic domain characterizing Φ^i. Therefore

$$-\ln P(\Phi^i \mid S^*(M^i)) \quad = \quad \begin{cases} 0 & \text{if } S^*(M^i) \in \mathcal{E}_{\Phi^i} \\ \infty & \text{otherwise} \end{cases} \tag{A.5}$$

On the other hand, the optimization algorithm in GEPPETO minimizes the cost function

$$\mathcal{F}_c(M^{1:2}) \;=\; |M^2 - M^1| \tag{A.6}$$

under the perceptual constraint

$$\mathcal{A}_c(M^i,\Phi^i) \;=\; \begin{cases} 0 & \text{if } S^*(M) \in \mathcal{E}_{\Phi^i} \\ \infty & \text{otherwise} \end{cases} \tag{A.7}$$

for $i \in \{1,2\}$. We already note that the form of these constraints are identical to the one approximated in Equation (A.5).

Optimization under constraints is performed in GEPPETO by gradient descent on the Lagrangian defined by

$$\mathcal{L}_G \;=\; \mathcal{F}_c + \mathcal{A}_c^1 + \mathcal{A}_c^2, \tag{A.8}$$

Hence, it appears that $\mathcal{L}_G$ is equal to $\mathcal{L}_B$ in equation (A.3) and therefore

$$P(M^{1:2} \mid \Phi^{1:2}\,[C_m = L]) \;\propto\; e^{-\mathcal{L}_G}. \tag{A.9}$$

This shows that finding a sequence of control variables $M^{1:2}$ that maximizes its posterior probability $P(M^{1:2} \mid \Phi^{1:2}\,[C_m = L])$ its equivalent to minimizing the corresponding cost function under the perceptual constraints defined by GEPPETO. This complete the proof that the optimal control approach performed by GEPPETO is included as a special case of the Bayesian model.

Appendix B

Derivation of planning process with continuous coherence variables

This section details the derivation of the motor planning process, provided in Eq (8.26) using the continuous coherence variables form:

$$P([C = c] \mid [S_M = s_1]\ [S_\Phi = s_2]) \quad \propto \quad 1 - c + 2c\ e^{-d(s_1, s_2)}, \tag{B.1}$$

The planning process is computed from the joint probability distribution as:

$$\begin{aligned} & P([M = m] \mid \Phi\ C_A\ C_S) \\ & \quad \propto \int P([M = m]\ S_M\ A_M\ \Phi\ S_\Phi\ A_\Phi\ C_S\ C_A)\ dS_M dS_\Phi dA_M dA_\Phi \\ & \quad \propto \int P(S_M \mid [M = m]) P(S_\Phi \mid \Phi) P(C_S \mid S_\Phi\ S_M)\ dS_M dS_\Phi \\ & \qquad \int P(A_M \mid [M = m]) P(A_\Phi \mid \Phi) P(C_A \mid A_\Phi\ A_M)\ dA_M dA_\Phi \end{aligned} \tag{B.2}$$

were terms $P(M)$ and $P(\Phi)$ were included in the proportionality symbol, since they are assumed to be constant. Since $P(S_M \mid [M = m])$ and $P(A_M \mid [M = m])$ are Dirac delta functions centered on $\rho_S(m)$ and $\rho_A(m)$, integration over S_M and A_M leads to:

$$\begin{aligned} & P([M = m] \mid \Phi\ C_A\ C_S) \\ & \quad \propto \int P(S_\Phi \mid \Phi) P(C_S \mid S_\Phi\ [S_M = \rho_S(m)])\ dS_\Phi \\ & \qquad \int P(A_\Phi \mid \Phi) P(C_A \mid A_\Phi\ [A_M = \rho_A(m)])\ dA_\Phi. \end{aligned} \tag{B.3}$$

Then, substituting $P(C_S \mid S_\Phi\ [S_M = \rho_S(m)])$ and $P(C_A \mid A_\Phi\ [A_M = \rho_A(m)])$ in Eq (B.3) with their expression given by Eq (B.1) gives:

$$\begin{aligned} & P([M = m] \mid \Phi\ [C_A = c_A]\ [C_S = c_S]) \\ & \quad \propto \int P(S_\Phi \mid \Phi) \left(1 - c_S + 2c_S e^{-d(S_\Phi, \rho_S(m))}\right)\ dS_\Phi \\ & \qquad \int P(A_\Phi \mid \Phi) \left(1 - c_A + 2c_A e^{-d(A_\Phi, \rho_A(m))}\right) dA_\Phi. \end{aligned} \tag{B.4}$$

Next, developing, rearranging terms, and using the fact that the integral of $P(S_\Phi \mid \Phi)$ and

$P(A_\Phi \mid \Phi)$ gives value 1 (they are probability distribution) gives:

$$\begin{aligned} & P([M=m] \mid \Phi\ [C_A=c_A]\ [C_S=c_S]) \\ \propto\ & (1-c_A)(1-c_S) \\ & +2(1-c_A)c_S \int P(S_\Phi \mid \Phi) e^{-d(S_\Phi,\rho_S(m))}\ dS_\Phi \\ & +2(1-c_S)c_A \int P(A_\Phi \mid \Phi) e^{-d(A_\Phi,\rho_A(m))}\ dA_\Phi \\ & +4c_Ac_S \int P(A_\Phi \mid \Phi) e^{-d(A_\Phi,\rho_A(m))} \int P(S_\Phi \mid \Phi) e^{-d(S_\Phi,\rho_S(m))}. \end{aligned} \tag{B.5}$$

And finally, by denoting

$$\Xi_A(\rho_A(m)) = \int P(A_\Phi \mid \Phi) e^{-d_A(A_\Phi,\rho_A(m))}\ dA_\Phi \tag{B.6}$$

$$\Xi_S(\rho_S(m)) = \int P(S_\Phi \mid \Phi) e^{-d_S(S_\Phi,\rho_S(m))}\ dS_\Phi, \tag{B.7}$$

Eq (B.5) becomes:

$$\begin{aligned} & P([M=m] \mid \Phi\ [C_A=c_A]\ [C_S=c_S]) \\ \propto\ & (1-c_A)(1-c_S) \\ & +2(1-c_A)c_S\Xi_S(\rho_S(m)) + 2(1-c_S)c_A\Xi_A(\rho_A(m)) \\ & +4c_Ac_S\Xi_S(\rho_S(m))\Xi_A(\rho_A(m)). \end{aligned} \tag{B.8}$$

Now we need to compute Eqs (B.6) and (B.7). Here is where the quadratic choice of measure $d(a,b) = \frac{(a-b)^2}{2\xi}$ becomes useful. Indeed, since $P(S_\Phi \mid \Phi)$ and $P(A_\Phi \mid \Phi)$ are multivariate normal distributions of parameters $(\mu_A^\Phi,\Sigma_A^\Phi)$ and $(\mu_S^\Phi,\Sigma_S^\Phi)$, Eqs (B.6) and (B.7) can be seen as (proportional to) the convolution product of two multivariate normal distributions:

$$\Xi_A(\rho_A(m)) = (2\pi)^{\frac{n_A}{2}}\xi_A \int \mathcal{N}(A_\Phi;\mu_A^\Phi,\Sigma_A^\Phi)\mathcal{N}(A_\Phi-\rho_A(m);0,\xi_A.I_{n_A})\ dA_\Phi \tag{B.9}$$

$$\Xi_S(\rho_S(m)) = (2\pi)^{\frac{n_S}{2}}\xi_S \int \mathcal{N}(S_\Phi;\mu_S^\Phi,\Sigma_S^\Phi)\mathcal{N}(S_\Phi-\rho_S(m);0,\xi_S.I_{n_S})\ dS_\Phi \tag{B.10}$$

where factors $(2\pi)^{\frac{n_A}{2}}\xi_A$ and $(2\pi)^{\frac{n_A}{2}}\xi_A$ come from the fact that the exponential terms in Eqs (B.6) and (B.7) are not normalized, n_A and n_S being the dimensionality of sensory spaces, and I_{n_A}, I_{n_S} are identity matrices or rank n_A and n_S. And knowing that the convolution of two multivariate normal distributions of parameters (μ_1,Σ_1) and (μ_1,Σ_1) is a multivariate normal distribution of parameters $(\mu_1+\mu_2,\Sigma_1+\Sigma_2)$ (refs (Bishop, 2006)), Eqs (B.9) and (B.10) become:

$$\Xi_A(\rho_A(m)) = (2\pi)^{\frac{n_A}{2}}\xi_A\mathcal{N}(\rho_A(m);\mu_A^\Phi,\Sigma_A^\Phi+\xi_AI_{n_A}) \tag{B.11}$$

$$\Xi_S(\rho_S(m)) = (2\pi)^{\frac{n_S}{2}}\xi_S\mathcal{N}(\rho_S(m);\mu_S^\Phi,\Sigma_S^\Phi+\xi_SI_{n_S}). \tag{B.12}$$

This ends the computation of the motor planning question:

$$\begin{aligned} & P([M=m] \mid \Phi\ [C_A=c_A]\ [C_S=c_S]) \\ \propto\ & (1-c_A)(1-c_S)(2\pi)^{-\frac{n_A+n_S}{2}} \\ & +2(1-c_A)c_S\xi_S(2\pi)^{\frac{-n_A}{2}}\mathcal{N}(\rho_S(m);\mu_S^\Phi,\Sigma_S^\Phi+\xi_AI_{n_A}) \\ & +2(1-c_S)c_A\xi_A(2\pi)^{\frac{-n_S}{2}}\mathcal{N}(\rho_A(m);\mu_A^\Phi,\Sigma_A^\Phi+\xi_AI_{n_A}) \\ & +4c_Ac_S\xi_S\xi_A\mathcal{N}(\rho_S(m);\mu_S^\Phi,\Sigma_S^\Phi+\xi_SI_{n_S})\mathcal{N}(\rho_A(m);\mu_A^\Phi,\Sigma_A^\Phi+\xi_AI_{n_A}). \end{aligned} \tag{B.13}$$

$$
\begin{aligned}
&P([M=m] \mid \Phi\ [C_A=c_A]\ [C_S=c_S]) \\
&\quad\propto\ (1-c_A)(1-c_S)(2\pi)^{-\frac{n_A+n_S}{2}} \\
&\qquad +2(1-c_A)c_S\xi_S(2\pi)^{\frac{-n_A}{2}}P_{\xi_S}(\rho_A(m)|\Phi)+2(1-c_S)c_A\xi_A(2\pi)^{\frac{-n_S}{2}}P_{\xi_A}(\rho_A(m)|\Phi) \\
&\qquad +4c_Ac_S\xi_S\xi_AP_{\xi_S}(\rho_A(m)|\Phi)P_{\xi_A}(\rho_A(m)|\Phi). \qquad \text{(B.14)}
\end{aligned}
$$

Appendix C

Illustration of results for all subjects

In this annex we provide additional figures illustrating results for all subjects in the experimental studies presented in Chapter 9.

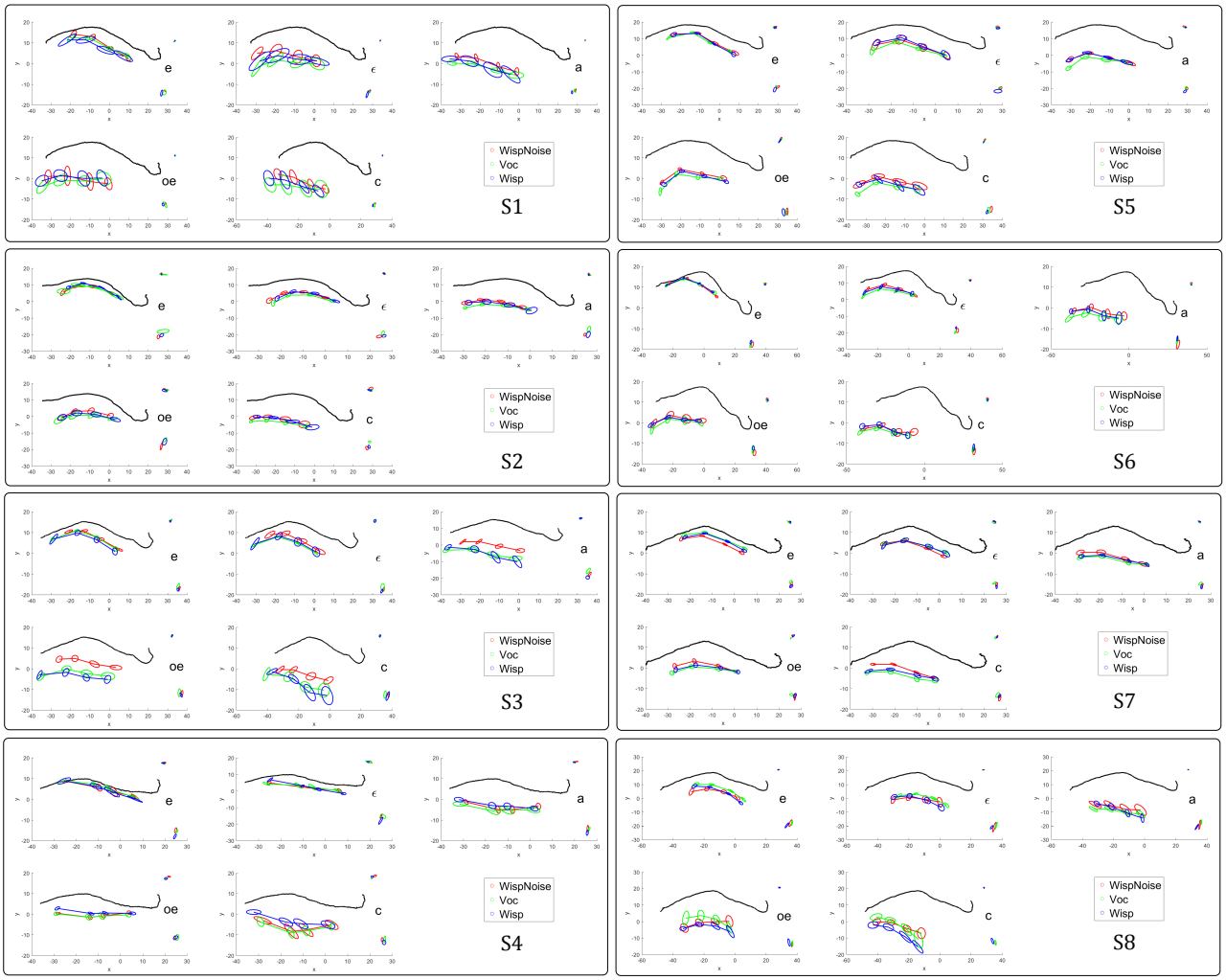

Figure C.1: Average productions and dispersion ellipses of each subjects, for each vowel in the three conditions

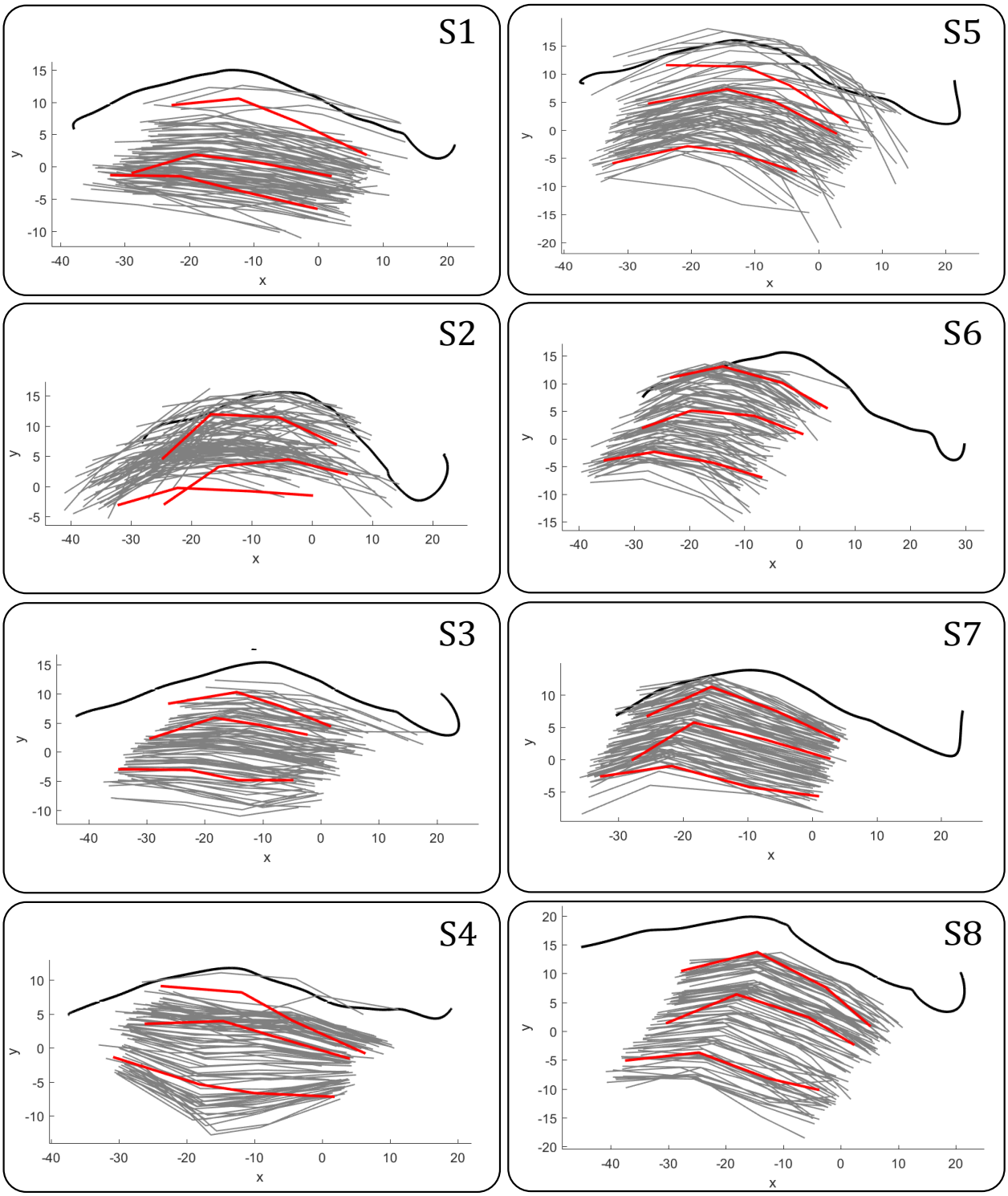

Figure C.2: All tongues postures attained during the reaching task for all subjects

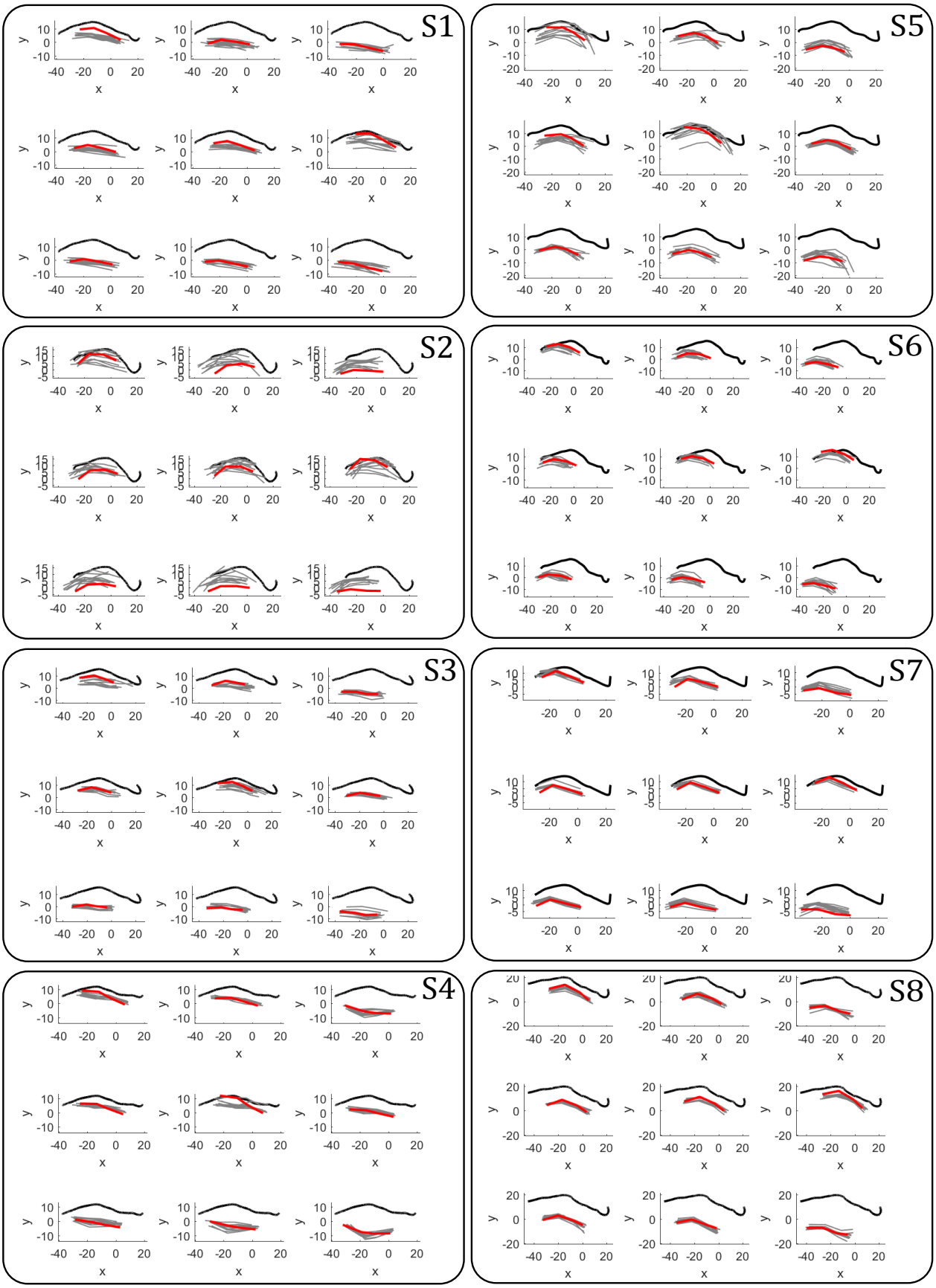

Figure C.3: Tongues postures attained for each target during the reaching task for all subjects

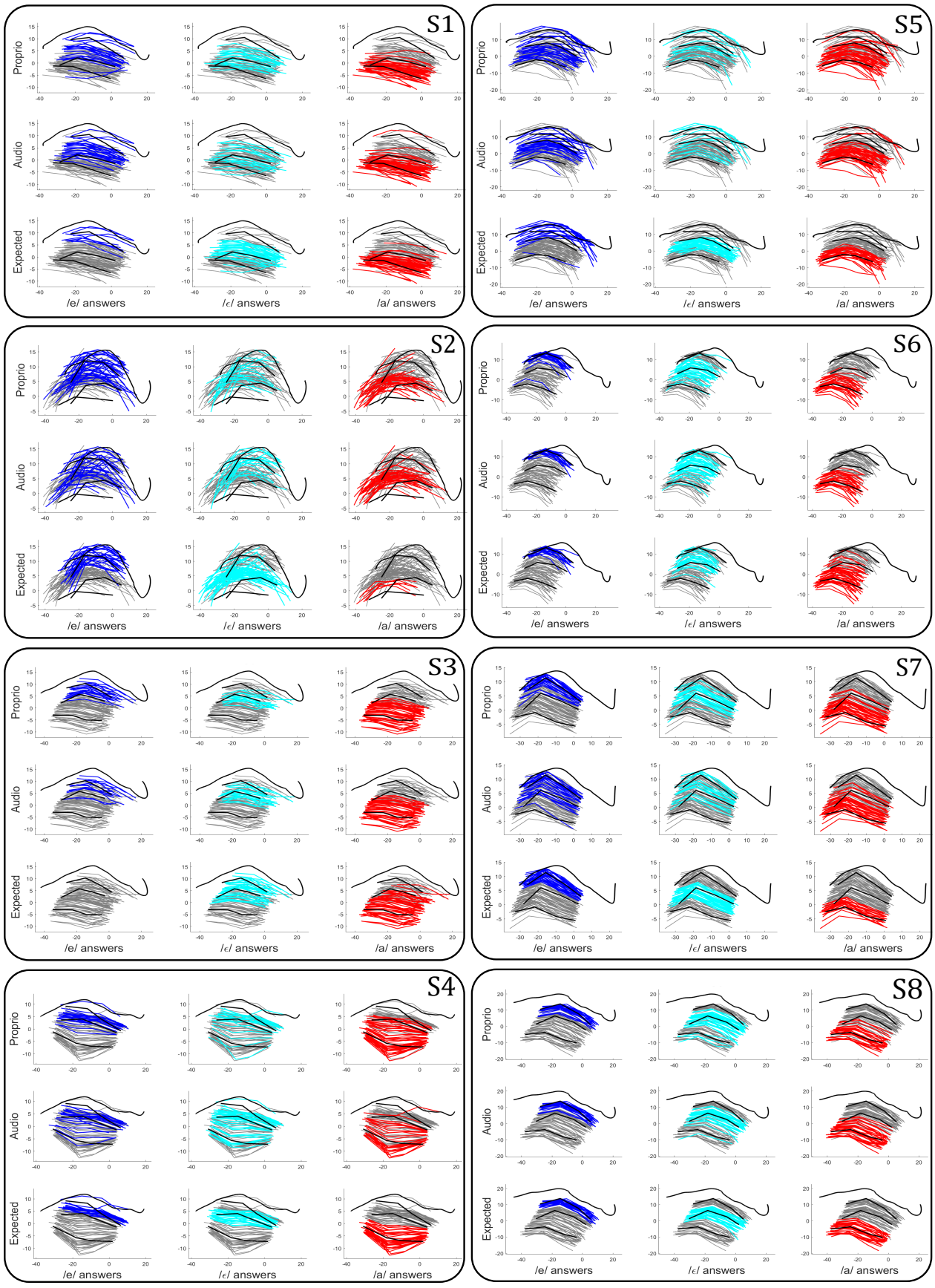

Figure C.4: Tongue postures labeled with expected answers (right panels) and subject's proprioceptive and auditory answers.

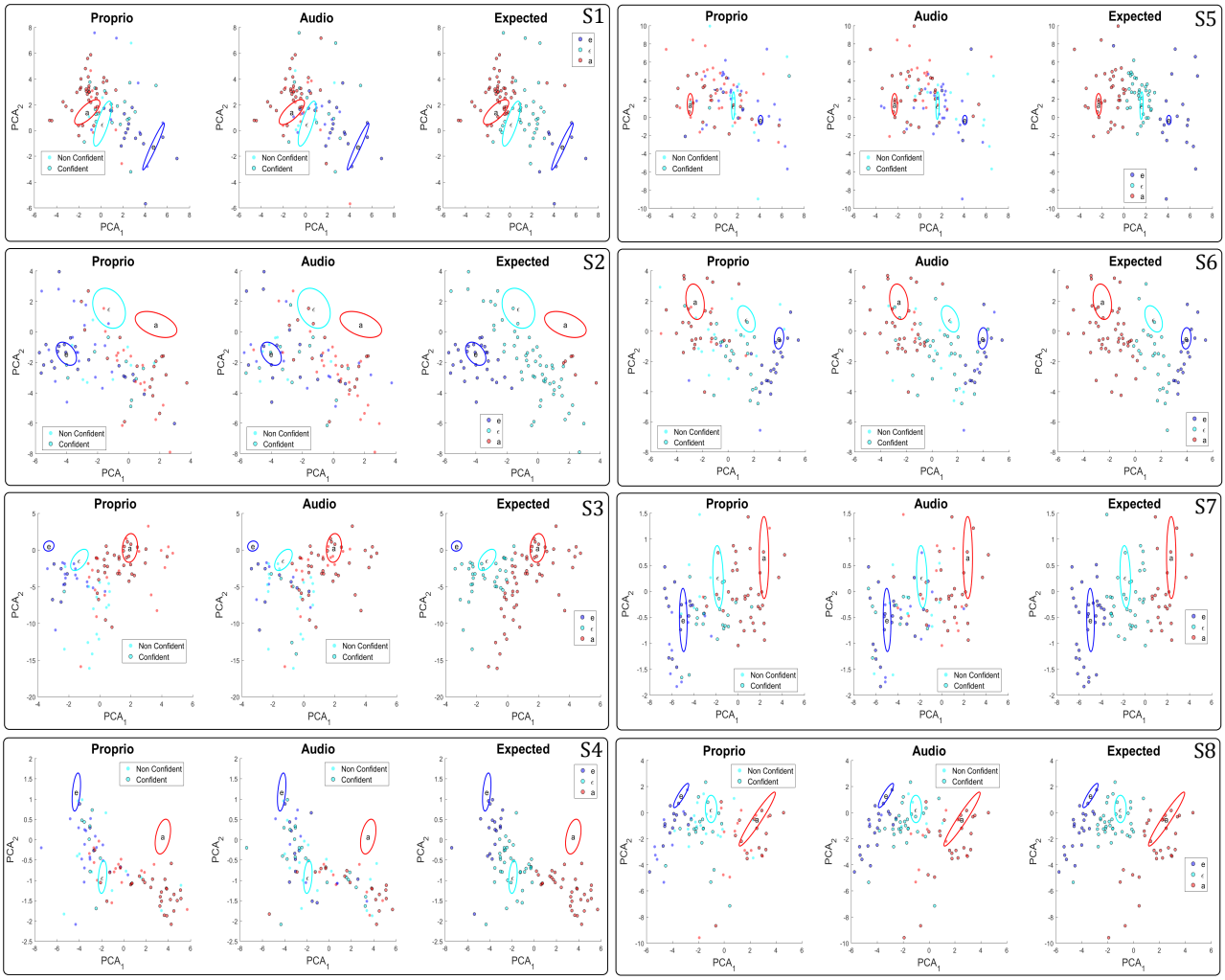

Figure C.5: Same as Fig C.4, but represented in PCA space

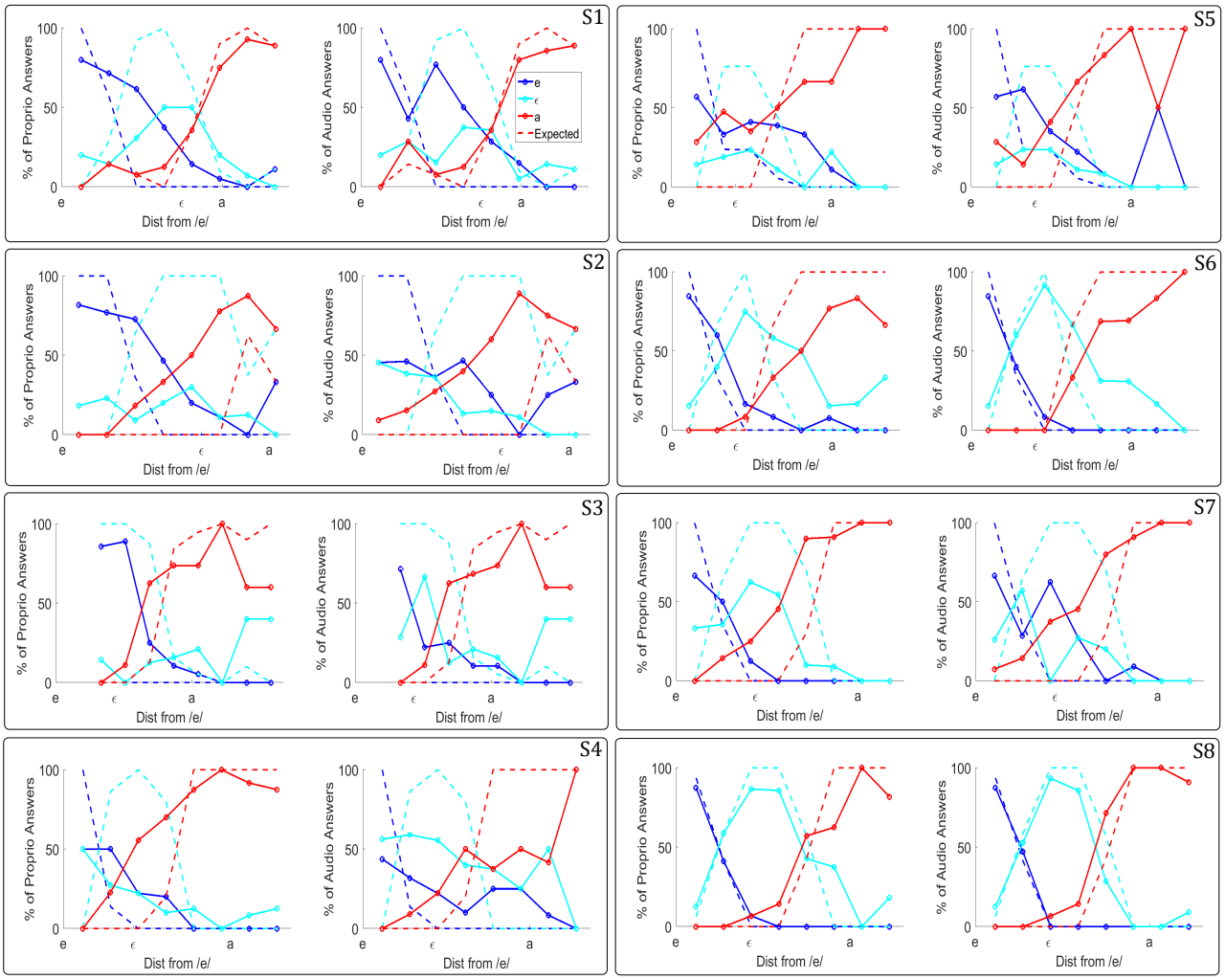

Figure C.6: Categorization curves as functions of the Mahalanobis distance from vowel /e/, for all subjects

Appendix D

Supporting Informations of Chapter 10

Supporting information S1.
Detailed model definition

This section provides additional technical description of the way our model is defined. The model definition applies the Bayesian Programming methodology [1], that proceeds in three steps: first, variables are selected and defined, second, the joint probability distribution over these variables is defined, usually by decomposing it as a product of probability distributions, which are simplified thanks to conditional independence hypotheses, and third and last, each of the terms in the decomposition is associated to a parametric form and a manner to identify these parameters (whether by experimental learning or by a priori definitions). We now provide the definition of the model, following each of these steps.

Variable definitions

For convenience, we recall here the variable definitions provided in the main text. Variables M, A_M, S_M, A_Φ and S_Φ are one dimensional continuous variables, i.e., each is in IR. Variable Φ is a three-valued categorical variable, with $\Phi = \{/i/, /\varepsilon/, /a/\}$. Finally, variables C_S and C_A are binary variables, i.e. two-valued categorical variables, with $C_S = C_A = \{0, 1\}$.

Decomposition of the joint probability distribution

With these variables, the joint probability distribution that defines our model mathematically is $P(M\ S_M\ A_M\ \Phi\ S_\Phi\ A_\Phi\ C_S\ C_A)$. Choosing a variable ordering and applying the chain rule, it is equal to:

$$\begin{aligned} & P(M\ S_M\ A_M\ \Phi\ S_\Phi\ A_\Phi\ C_S\ C_A) \\ = \ & P(M)P(A_M \mid M)P(S_M \mid A_M\ M) \\ & P(\Phi \mid S_M\ A_M\ M)P(A_\Phi \mid \Phi\ S_M\ A_M\ M)P(S_\Phi \mid A_\Phi\ \Phi\ S_M\ A_M\ M) \\ & P(C_A \mid S_\Phi\ A_\Phi\ \Phi\ S_M\ A_M\ M)P(C_S \mid C_A\ S_\Phi\ A_\Phi\ \Phi\ S_M\ A_M\ M)\ . \end{aligned} \tag{1}$$

We now apply conditional independence hypotheses to simplify some of these terms.

The first two, $P(M)$ and $P(A_M \mid M)$, are left unchanged. The term $P(S_M \mid A_M\ M)$ is simplified into $P(S_M \mid M)$: this assumes that the cognitive agent's knowledge about the somatosensory consequence of some motor command m is independent of the acoustic consequence of m when m is known. In other words, the main cause of somatosensory signals S_M is assumed to be motor commands, and the cognitive agent dismisses the additional information carried out by A_M about S_M. What is lost in this approximation is the possible physical effect of acoustic waves provoked by sound production on somatosensory sensors; an effect likely to be negligible. It has to be noted that this conditional independence hypothesis between S_M and A_M given M does not entail at all independence between S_M and A_M. For instance, in the model, the cognitive agent can retrieve $P(S_M\ A_M) \propto \sum_M P(M)P(A_M \mid M)P(S_M \mid M)$ which is not equal to $P(S_M)P(A_M)$ in the general case. This means that the model contains knowledge about relations between auditory and somatosensory consequences of motor commands, but it does not store it as an explicit piece of knowledge.

The three next terms are assumed to constitute a separate piece of model, independent from knowledge about motor commands and their sensory consequences, so that variables S_M, A_M and M can be dropped. This yields $P(\Phi)$, $P(A_\Phi \mid \Phi)$ and $P(S_\Phi \mid A_\Phi\ \Phi)$. Furthermore, using a similar conditional independence hypothesis as above, we assume that phonemes are characterized independently into

acoustic and somatosensory spaces. In other words, the somatosensory characterization S_Φ of some phoneme ϕ is supposed to be independent of the acoustic characterization A_Φ of this phoneme, when ϕ is known. Therefore, $P(S_\Phi \mid A_\Phi\ \Phi)$ is simplified into $P(S_\Phi \mid \Phi)$.

Finally, the last two terms concerns coherence variables, which we, as modelers, connect explicitly to chosen variables: first, variable C_A serves as a connector between auditory representations A_M and A_Φ, second, variable C_S serves as a connector between somatosensory representations S_M and S_Φ. This yields terms $P(C_A \mid A_M\ A_\Phi)$ and $P(C_S \mid S_M\ S_\Phi)$.

Replacing each term of Eq (1) by its simplified form, we obtain the decomposition of the joint distribution shown in the main text (Eq. (1)), which we repeat here:

$$\begin{aligned} & P(M\ S_M\ A_M\ \Phi\ S_\Phi\ A_\Phi\ C_S\ C_A) \\ = \quad & P(M)P(A_M \mid M)P(S_M \mid M) \\ & P(\Phi)P(A_\Phi \mid \Phi)P(S_\Phi \mid \Phi) \\ & P(C_A \mid A_M\ A_\Phi)P(C_S \mid S_M\ S_\Phi)\ . \end{aligned} \tag{2}$$

Parametric forms

Parametric forms for all terms in Eq (2) are provided in the main text. We still describe here, in a bit more detail, the properties of coherence variables (demonstrations are available elsewhere [2, 1]. Recall that coherence variables are binary variables associated with Dirac distributions that enforce matching constraints. Consider C_A: we have defined

$$P\left([C_A = 1] \mid [A_M = a_m]\ [A_\Phi = a_\phi]\right) \quad \coloneqq \quad \begin{cases} 1 & \text{if } a_m = a_\phi; \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Given this definition, coherence variables can be used, during inference, as "switches", allowing the modeler to control the propagation of information throughout the model. There are three cases to consider.

First, when left unspecified in the computed question, the switch is open, and portions of the model on each side of the coherence variable do not exchange information. For instance, in the model, the portion of the model about phoneme characterizations can be "separated" from the portion about the sensory consequences of motor commands: $P(M \mid A_M\ S_M)$ can be completely computed by involving terms $P(M)$, $P(A_M \mid M)$ and $P(S_M \mid M)$, as other terms of the decomposition are "beyond" the coherence variables, which are not specified in $P(M \mid A_M\ S_M)$. Computing the motor cause of some sensed sensory event A_M, S_M would only involve knowledge about the way motor commands provoke sensory effects; whatever the phonological plausibility of this sensory event.

Second, when set to 1 in the computed question, the switch is closed, and variables on each side of the coherence variable are forced to have equal values, so that information about one variable propagates and constrains the other. For instance, in the model, computing $P(M \mid A_M\ S_M\ [C_A = 1]\ [C_S = 1])$ would be influenced by phonological knowledge, as the "switches" are here closed so that A_M is constrained by A_Φ and S_M is constrained by S_Φ. Here, computing the motor cause of some sensed sensory event, A_M and S_M, would also involve the phonological plausibility of this sensory event.

Third and finally, when set to 0, variables are forced to be different, which is less useful in practice – but see [1, p 139].

References

1. Bessière P, Mazer E, Ahuactzin JM, Mekhnacha K. Bayesian Programming. Boca Raton, Florida: CRC Press; 2013.

2. Gilet E, Diard J, Bessière P. Bayesian Action–Perception Computational Model: Interaction of Production and Recognition of Cursive Letters. PLOS ONE. 2011;6(6):e20387.

Supporting information S2.
Derivation of Bayesian inference equations

Here we detail to derivation of Eqs (8) to (13) of the main text. All computations rest on the same principle, i.e. the application of Bayesian inference. Bayes rule dictates how to compute conditional probabilities when the joint probability distribution of variables is known. For instance, in the case of two probabilistic variables A and B, the conditional probability $P(A \mid B)$ is computed from the joint probability $P(A\ B)$ as (provided that $P(B) \neq 0$):

$$P(A \mid B) = \frac{P(A\ B)}{P(B)} = \frac{P(A\ B)}{\sum\limits_{A} P(A\ B)}.$$

This is generalized to $(n+1)$ variables, $\{A_1, A_2, \cdots, A_n, B\}$ as:

$$P(A_1 \mid B) = \frac{P(A_1\ B)}{P(B)} = \frac{\sum\limits_{/\{A_1,B\}} P(A_1\ \cdots A_n\ B)}{\sum\limits_{/\{B\}} P(A_1\ \cdots A_n\ B)}, \tag{1}$$

where $\sum\limits_{/X}$ denotes summing over all variables but those included in the set X.

We will also consider conditional probabilities of more than two variables, which are obtained in a similar way:

$$P(A_1 \mid A_2\ B) = \frac{P(A_1\ A_2\ B)}{P(A_2\ B)} = \frac{\sum\limits_{/\{A_1,A_2,B\}} P(A_1 \cdots A_n\ B)}{\sum\limits_{/\{A_2,B\}} P(A_1 \cdots A_n\ B)}, \tag{2}$$

$$P(A_1 \mid A_2\ A_3\ B) = \frac{P(A_1\ A_2\ A_3\ B)}{P(A_2\ A_3\ B)} = \frac{\sum\limits_{/\{A_1,A_2,A_3,B\}} P(A_1 \cdots A_n\ B)}{\sum\limits_{/\{A_2,A_3,B\}} P(A_1 \cdots A_n\ B)}. \tag{3}$$

Independence of M and Φ.

The first result to demonstrate concerns the conditional independence between M and Φ. We have to show that the model predicts $P(M \mid \Phi) = P(M)$. This result is obtained by computing $P(M \mid \Phi)$ from the joint probability distribution $P(M\ S_M\ A_M\ \Phi\ S_\Phi\ A_\Phi\ C_S\ C_A)$, using Bayes rule in Eq (1) :

$$P(M \mid \Phi) = \frac{P(M\ \Phi)}{P(\Phi)} = \frac{\sum\limits_{/\{M,\Phi\}} P(M\ S_M\ A_M\ \Phi\ S_\Phi\ A_\Phi\ C_S\ C_A)}{\sum\limits_{/\{\Phi\}} P(M\ S_M\ A_M\ \Phi\ S_\Phi\ A_\Phi\ C_S\ C_A)}. \tag{4}$$

In order to avoid confusions, we draw attention to the notation that is employed here. The domain of the joint probability distribution is composed of discrete and continuous variables. Usually, one writes P for probability distributions over discrete variables and p for probability densities over continuous variables. For simplicity, we chose not to make this distinction here. Similarly, all summations and integrals are denoted by the sign $\sum$, even when rigorously it is the $\int$ sign that should be used for continous variables.

The summations in Eq (4) can be performed by replacing the joint probability distribution by its decomposition:

$$\begin{aligned} P(M\ S_M\ A_M\ \Phi\ S_\Phi\ A_\Phi\ C_S\ C_A) &= P(M)P(A_M \mid M)P(S_M \mid M) \\ & \quad P(\Phi)P(A_\Phi \mid \Phi)P(S_\Phi \mid \Phi) \\ & \quad P(C_A \mid A_M\ A_\Phi)P(C_S \mid S_M\ S_\Phi). \end{aligned} \tag{5}$$

Since probability distributions are normalized, reorganizing the sums in the numerator leads to unit factors, and only the product $P(M)P(\Phi)$ remains. The same holds for the denominator, where in addition the sum over M also reduces the factor $P(M)$ to 1, and hence only $P(\Phi)$ remains. Altogether, Eq (4) becomes:

$$P(M \mid \Phi) = \frac{P(M)P(\Phi)}{P(\Phi)} = P(M), \tag{6}$$

proving that M is independent of Φ.

Similar steps enable to prove that variables Φ and A_M are also independent.

Production questions.

Three production questions were defined in the text. They corresponded to the planning of motor commands M for the production of phoneme Φ, considering that either C_A, C_S, or both, are set to 1.

The first and second questions, $P(M \mid \Phi\ [C_A = 1])$ and $P(M \mid \Phi\ [C_S = 1])$, are computed in the same way using Eq (2). We only detail here the computations in the case of $P(M \mid \Phi\ [C_A = 1])$. From Eq (2) we have:

$$P(M \mid \Phi\ [C_A = 1]) \quad \propto \sum_{/\{M,\Phi,C_A\}} P(M\ S_M\ A_M\ \Phi\ S_\Phi\ A_\Phi\ C_S\ [C_A = 1]), \tag{7}$$

in which the denominator, independent of M, has been included in the proportionality symbol $\propto$. The summation in Eq (7) is performed using the decomposition of the joint probability distribution in Eq (1) in the text. Again, most terms sum to 1, such that Eq (7) results in:

$$P(M \mid \Phi\ [C_A = 1]) \quad \propto \sum_{A_M, A_\Phi} P(A_M \mid M)P(A_\Phi \mid \Phi)P([C_A = 1] \mid A_\Phi\ A_M), \tag{8}$$

where $P(M)$ and $P(\Phi)$ were also included in the proportionality symbol since they are assumed to be uniform. $P(A_M \mid M)$ in Eq (8) is a Dirac delta function given by:

$$P([A_M = a] \mid [M = m]) \quad \coloneqq \quad \delta\left(a - \rho_A(m)\right), \tag{9}$$

such that, in the sum over values taken by A_M in Eq (8), only the term involving value $a = \rho_A(m)$ does not vanish. Eq (8) becomes:

$$P([M = m] \mid \Phi\ [C_A = 1]) \quad \propto \sum_{A_\Phi} P(A_\Phi \mid \Phi)P([C_A = 1] \mid A_\Phi\ [A_M = \rho_A(m)]). \tag{10}$$

The second factor in Eq (10) corresponds to the sensory-matching constraint defined as:

$$P\left([C_A = 1] \mid [A_M = a_m]\ [A_\Phi = a_\phi]\right) \quad \coloneqq \quad \begin{cases} 1 & \text{if } a_m = a_\phi; \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Hence, in the sum over values taken by A_Φ in Eq (10), only the term involving value $\rho_A(m)$ does not vanish. Hence Eq (10) becomes:

$$\boxed{P([M = m] \mid \Phi\ [C_A = 1]) \propto P([A_\Phi = \rho_A(m)] \mid \Phi).} \tag{12}$$

The third question, $P(M \mid \Phi\ [C_A = 1][C_S = 1])$, is computed from Eq (3):

$$P(M \mid \Phi\ [C_A = 1]\ [C_S = 1])$$
$$\propto \sum_{/\{M,\Phi,C_A,C_S\}} P(M\ S_M\ A_M\ \Phi\ S_\Phi\ A_\Phi\ [C_S = 1]\ [C_A = 1]), \tag{13}$$

in which again, the denominator, being independent of M, has been included in the proportionality symbol $\propto$. As before, the summation in Eq (13) is performed using the decomposition of the joint probability distribution in Eq (5), resulting in:

$$P(M \mid \Phi\ [C_A = 1]\ [C_S = 1])$$
$$\propto \sum_{\{A_M,A_\Phi\}} P(A_M \mid M)P(A_\Phi \mid \Phi)P([C_A = 1] \mid A_\Phi\ A_M)$$
$$\sum_{\{S_M,S\Phi\}} P(S_M \mid M)P(S_\Phi \mid \Phi)P([C_S = 1] \mid S_\Phi\ S_M) \tag{14}$$

where again $P(M)$ and $P(\Phi)$ were included in the proportionality symbol since they are assumed to be uniform.

As before, in the sum over values taken by A_M and S_M in Eq (14), only the terms involving values $a = \rho_A(m)$ and $s = \rho_S(m)$ do not vanish. Hence Eq (14) becomes:

$$P([M = m] \mid \Phi\ [C_A = 1]\ [C_S = 1])$$
$$\propto \sum_{A_\Phi} P(A_\Phi \mid \Phi)P([C_A = 1] \mid A_\Phi\ [A_M = \rho_A(m)])$$
$$\sum_{S_\Phi} P(S_\Phi \mid \Phi)P([C_S = 1] \mid S_\Phi\ [S_M = \rho_S(m)]). \tag{15}$$

Then again, with the definition of the two sensory matching constraints, in the sum over values taken by A_Φ and S_Φ in Eq (15), only terms involving values $\rho_A(m)$ $\rho_S(m)$ do not vanish. Hence Eq (15) becomes:

$$\boxed{P([M = m] \mid \Phi\ [C_A = 1]\ [C_S = 1]) \propto P([A_\Phi = \rho_A(m)] \mid \Phi)P([S_\Phi = \rho_S(m)] \mid \Phi).} \tag{16}$$

Perception questions.

Three perception questions were defined in the text. They corresponded to the inference of phoneme identity Φ given the auditory input A_M, considering that either C_A, C_S, or both, are set to 1.

The first question, $P(\Phi \mid A_M\ [C_A = 1])$ is computed from Eq (2):

$$P(\Phi \mid A_M\ [C_A = 1]) \quad = \quad \frac{\sum_{/\{A_M,\Phi,C_A\}} P(M\ S_M\ A_M\ \Phi\ S_\Phi\ A_\Phi\ C_S\ [C_A = 1])}{\sum_{/\{A_M,C_A\}} P(M\ S_M\ A_M\ \Phi\ S_\Phi\ A_\Phi\ C_S\ [C_A = 1])}. \tag{17}$$

The decomposition of the joint probability distribution (Eq (5)) is used to perform the sums in Eq (17). Sums over C_S, S_M and S_Φ reduce to factors of 1, such that:

$$P(\Phi \mid [A_M = a]\ [C_A = 1])$$
$$= \frac{\sum_{M,A_\Phi} P([A_M = a] \mid M)P(A_\Phi \mid \Phi)P([C_A = 1] \mid A_\Phi\ [A_M = a])}{\sum_{\Phi,M,A_\Phi} P([A_M = a] \mid M)P(A_\Phi \mid \Phi)P([C_A = 1] \mid A_\Phi\ [A_M = a])}, \tag{18}$$

were terms $P(M)$ and $P(\Phi)$ are taken outside of the summations, since they are assumed to be constant, and further simplified since they appear both in the numerator and the denominator.

Performing the sums over A_Φ and reorganizing terms leads to the following steps:

$$\begin{aligned} P(\Phi \mid [A_M = a]\ [C_A = 1]) &= \frac{\sum\limits_{M} P([A_M = a] \mid M) P([A_\Phi = a] \mid \Phi)}{\sum\limits_{\Phi, M} P([A_M = a] \mid M) P([A_\Phi = a] \mid \Phi)}, \\ &= \frac{P([A_\Phi = a]|\Phi) \sum\limits_{M} P([A_M = a] \mid M)}{\sum\limits_{\Phi} P([A_\Phi = a]|\Phi) \sum\limits_{M} P([A_M = a] \mid M)}. \end{aligned} \tag{19}$$

Finally, simplifying terms on the numerator and denominator of Eq (19) leads to:

$$\boxed{P(\Phi \mid [A_M = a]\ [C_A = 1]) = \frac{P([A_\Phi = a] \mid \Phi)}{\sum\limits_{\Phi} P([A_\Phi = a] \mid \Phi)}.} \tag{20}$$

The second perception question, $P(\Phi \mid A_M\ [C_S = 1])$, is also computed from Eq (2).

$$P(\Phi \mid A_M\ [C_S = 1]) = \frac{\sum\limits_{/\{A_M, \Phi, C_S\}} P(M\ S_M\ A_M\ \Phi\ S_\Phi\ A_\Phi\ C_A\ [C_S = 1])}{\sum\limits_{/\{A_M, C_S\}} P(M\ S_M\ A_M\ \Phi\ S_\Phi\ A_\Phi\ C_A\ [C_S = 1])}. \tag{21}$$

Again, using the decomposition of the joint probability distribution enables to perform the sums in Eq (21). The sums over A_Φ and C_A reduce to 1 such that:

$$\begin{aligned} &P(\Phi \mid [A_M = a]\ [C_S = 1]) \\ &= \frac{\sum\limits_{M, S_\Phi, S_M} P([A_M = a] \mid M) P(S_M \mid M) P(S_\Phi \mid \Phi) P([C_S = 1] \mid S_\Phi\ S_M)}{\sum\limits_{\Phi, M, S_\Phi, S_M} P([A_M = a] \mid M) P(S_M \mid M) P(S_\Phi \mid \Phi) P([C_S = 1] \mid S_\Phi\ S_M)}, \end{aligned} \tag{22}$$

were terms $P(M)$ and $P(\Phi)$ were again taken outside of the sums, since they are assumed to be constant, and further simplified.

In the sums over M in Eq (22), the factor $P([A_M = a] \mid M)$ is zero unless M takes a value for which its image through the auditory-motor mapping ρ_A is a. In other words, the sum is reduced to the set of inverse images of a, $m_a \in \{\rho_A^{-1}(a)\}$, and Eq (22) becomes :

$$\begin{aligned} &P(\Phi \mid [A_M = a]\ [C_S = 1]) \\ &= \frac{\sum\limits_{m_a \in \{\rho_A^{-1}(a)\}} \sum\limits_{S_\Phi, S_M} P(S_M \mid [M = m_a]) P(S_\Phi \mid \Phi) P([C_S = 1] \mid S_\Phi\ S_M)}{\sum\limits_{m_a \in \{\rho_A^{-1}(a)\}} \sum\limits_{\Phi, S_\Phi, S_M} P(S_M \mid [M = m_a]) P(S_\Phi \mid \Phi) P([C_S = 1] \mid S_\Phi\ S_M)}. \end{aligned} \tag{23}$$

Next, in the sums over S_M, because of the Dirac delta function $P(S_M|[M = m_a])$, only the term $S_M = \rho_S(m_a)$ remains, hence Eq (23) becomes:

$$\begin{aligned} &P(\Phi \mid [A_M = a]\ [C_S = 1]) \\ &= \frac{\sum\limits_{m_a \in \{\rho_A^{-1}(a)\}} \sum\limits_{S_\Phi} P(S_\Phi \mid \Phi) P([C_S = 1] \mid S_\Phi\ [S_M = \rho_S(m_a)])}{\sum\limits_{m_a \in \{\rho_A^{-1}(a)\}} \sum\limits_{\Phi, S_\Phi} P(S_\Phi \mid \Phi) P([C_S = 1] \mid S_\Phi\ [S_M = \rho_S(m_a)])}. \end{aligned} \tag{24}$$

Finally, performing the sums over S_Φ, because of the sensory matching constraint $P([C_S = 1] \mid S_\Phi\ [S_M = \rho_S(m_a)])$, we obtain:

$$P(\Phi \mid [A_M = a]\ [C_S = 1]) = \frac{\sum\limits_{m_a \in \{\rho_A^{-1}(a)\}} P([S_\Phi = \rho_S(m_a)] \mid \Phi)}{\sum\limits_{m_a \in \{\rho_A^{-1}(a)\}} \sum\limits_{\Phi} P([S_\Phi = \rho_S(m_a)] \mid \Phi)}. \tag{25}$$

For an injective auditory-motor mapping ρ_A, as it is the case in normal conditions or in the global update hypothesis (linear mapping), auditory values a have a unique inverse image which is directly found as $a_m = \rho_A^{-1}(a)$. In this case the sum is reduced to a unique term and Eq (25) becomes:

$$\boxed{P(\Phi \mid [A_M = a]\ [C_S = 1]) = \frac{P([S_\Phi = \rho_S \circ \rho_A^{-1}(a)] \mid \Phi)}{\sum\limits_{\Phi} P([S_\Phi = \rho_S \circ \rho_A^{-1}(a)] \mid \Phi)}.} \tag{26}$$

In the local update hypothesis, the auditory-motor mapping is no longer injective. In Supplementary information S3 we indicate how the different inverse images are found.

The third perception question, $P(\Phi \mid A_M\ [C_A = 1][C_S = 1])$ is computed from Eq (3). The steps are similar to the previous cases and lead to:

$$\begin{aligned} &P(\Phi \mid [A_M = a]\ [C_S = 1]) \\ &= \frac{P([A_\Phi = a] \mid \Phi) \sum\limits_{m_a \in \{\rho_A^{-1}(a)\}} P([S_\Phi = \rho_S(m_a)] \mid \Phi)}{\sum\limits_{m_a \in \{\rho_A^{-1}(a)\}} \sum\limits_{\Phi} P([A_\Phi = a] \mid \Phi) P([S_\Phi = \rho_S(m_a)] \mid \Phi)}, \end{aligned} \tag{27}$$

which again, for injective auditory-motor mappings ρ_A, as in normal conditions or in the global update hypothesis, results in:

$$\boxed{P(\Phi \mid [A_M = a]\ [C_S = 1]) = \frac{P([A_\Phi = a] \mid \Phi) P([S_\Phi = \rho_S \circ \rho_A^{-1}(a)] \mid \Phi)}{\sum\limits_{\Phi} P([A_\Phi = a] \mid \Phi) P([S_\Phi = \rho_S \circ \rho_A^{-1}(a)] \mid \Phi)}.} \tag{28}$$

Supporting information S3.
Specification of parameters for the local update of the auditory-motor mapping ρ_A

In the local update hypothesis, the linear relation assumed for the auditory-motor mapping in normal condition must be abandoned. For computational simplicity, we assume that the new auditory-motor internal model is linear in a piece-wise manner, as illustrated in Fig 1 and given by:

$$\rho_A^{(u)}(m) \coloneqq \begin{cases} \rho_1(m) = m + \delta_A & \text{if } m \in [\eta - \omega \ ; \ \eta + \omega], \\ \rho_2(m) = \frac{\chi+\delta_A}{\chi} m + \delta_A \left(1 - \frac{\eta-\omega}{\chi}\right) & \text{if } m \in [\eta - \omega - \chi \ ; \ \eta - \omega), \\ \rho_3(m) = \frac{\chi-\delta_A}{\chi} m + \delta_A \left(1 + \frac{\eta+\omega}{\chi}\right) & \text{if } m \in \ (\eta + \omega \ ; \ \eta + \omega + \chi], \\ \rho_4(m) = m & \text{otherwise.} \end{cases} \tag{1}$$

Parameter δ_A corresponds to the magnitude of the update (equal to the perturbation in the case where full compensation is assumed). Parameters η and ω correspond respectively to the center and width of the updated region in the motor command domain. Parameter χ specifies the width in the motor command domain of the intermediate intervals joining the updated and non-updated intervals of the mapping. Note that the general update is recovered when $\chi \to +\infty$.

Eq (1) and Fig 1 allow to identify the inverse images of an auditory output a. Depending on the sign of $\delta_A - \chi$ and on the location of a there is one, two or three inverse images (we leave aside the case where $\delta_A = \chi$ for which there is an infinite number of inverse images corresponding to the segment $[\eta + \omega; \eta + \omega + \chi]$):

If $\delta_A > \chi$ and $a \in [(\eta + \omega + \chi) \ ; \ (\eta + \omega + \delta_A)]$, there are three inverse images, corresponding to:

$$\begin{cases} m_1 = \begin{cases} \rho_2^{-1}(a) = \frac{\chi}{\chi+\delta_A}\left(a - \delta_A\left(1 - \frac{\eta-\omega}{\chi}\right)\right) & \text{if } a \in [(\eta + \omega + \chi) \ ; \ (\eta - \omega + \delta_A)], \\ \rho_1^{-1}(a) = a - \delta_A & \text{otherwise,} \end{cases} \\ m_2 = \rho_3^{-1}(a) = \frac{\chi}{\chi-\delta_A}\left(a - \delta_A\left(1 + \frac{\eta+\omega}{\chi}\right)\right), \\ m_3 = \rho_4^{-1}(a) = a. \end{cases} \tag{2}$$

If $\delta_A > \chi$ and $a = \eta + \omega + \chi$, there are two inverse images, corresponding to:

$$\begin{cases} m_1 = \begin{cases} \rho_2^{-1}(a) = \frac{\chi}{\chi+\delta_A}\left(a - \delta_A\left(1 - \frac{\eta-\omega}{\chi}\right)\right) & \text{if } (\eta + \omega + \chi) > (\eta - \omega + \delta_A), \\ \rho_1^{-1}(a) = a - \delta_A & \text{otherwise,} \end{cases} \\ m_2 = \rho_4^{-1}(a) = \eta + \omega + \chi. \end{cases} \tag{3}$$

If $\delta_A < \chi$ or $a \notin [(\eta + \omega + \chi) \ ; \ (\eta + \omega + \delta_A)]$, there is a single inverse image corresponding to:

$$m = \begin{cases} \rho_1^{-1}(a) = a - \delta_A & \text{if } a \in \ [(\eta - \omega + \delta_A) \ ; \ (\eta + \omega + \chi)], \\ \rho_2^{-1}(a) = \frac{\chi}{\chi+\delta_A}\left(a - \delta_A\left(1 - \frac{\eta-\omega}{\chi}\right)\right) & \text{if } a \in \ [(\eta - \omega - \chi) \ ; \\ & \quad \min\left((\eta + \omega + \chi), (\eta - \omega + \delta_A)\right)], \\ \rho_4^{-1}(a) = a & \text{otherwise.} \end{cases} \tag{4}$$

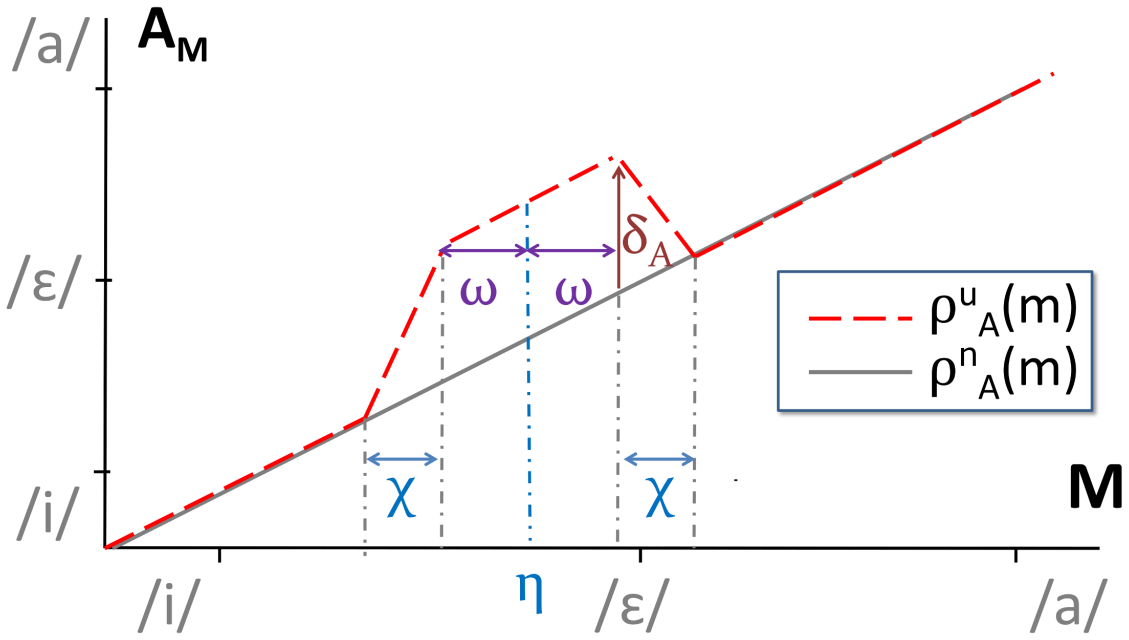

Figure 1: Auditory-motor mapping ρ_A before and after a local update. The plain gray straight line corresponds to $\rho_A^{(n)}$, the linear auditory-motor mapping in normal condition. The red dashed line corresponds to $\rho_A^{(u)}$, the locally updated auditory-motor mapping. Parameters η and ω correspond to the center and width, in the motor command domain, of the updated portion of the internal model. Parameter χ specifies the width, in the motor command domain, of the intermediate interval joining the updated and non-updated intervals of the mapping.

Supporting information S4.
Specification of parameters of the sensory characterizations of phonemes $P(A_\Phi \mid \Phi)$ and $P(S_\Phi \mid \Phi)$

S4.1. Normal condition.

The sensory characterization of phonemes are characterized by parameters $(\mu_A^\Phi, \sigma_A^\phi)$ and $(\mu_S^\Phi, \sigma_S^\phi)$. Phonemes are assumed to be equally spaced in both sensory spaces. In addition, units of auditory and somatosensory spaces are defined such that the distance between two neighboring phonemes is 1. Hence, in normal conditions the center of each sensory characterization is defined by:

$$\mu_A^{i\,(n)} = \mu_S^{i\,(n)} = 1, \tag{1}$$

$$\mu_A^{\varepsilon\,(n)} = \mu_S^{\varepsilon\,(n)} = 2, \tag{2}$$

$$\mu_A^{a\,(n)} = \mu_S^{a\,(n)} = 3. \tag{3}$$

In addition, in normal conditions all sensory characterizations are assumed to have same standard deviation, equal to $\frac{1}{8}$ of the distance between neighboring phonemes. In other words:

$$\sigma_S^{\phi\,(n)} = \sigma_A^{\phi\,(n)} = \frac{1}{8}, \qquad \phi \in \{\ \mathrm{i},\ \varepsilon,\ \mathrm{a}\ \}. \tag{4}$$

S4.2. Adapted condition: relation between parameters μ_A^ε and σ_A^ε.

In the text we suggested that choosing a correct combination of parameters μ_A and σ_A could reproduce the observation reported in L-14 by canceling out the boundary shift induced by each of the parameter updates on one side of the auditory space. Here we derive the relation that was used in simulations involving this combined update of μ_A and σ_A.

In the case of a perturbation of vowel /ε/ toward vowel /a/, the boundary shift was reported to be significant in the portion of auditory space between vowel /i/ and /ε/ and not significant in the portion of auditory space between vowels /ε/ and /a/. In order to reproduce this observation, we need to find a relation between parameters μ_A^ε and σ_A^ε that leaves unchanged the location of the boundary between vowels /ε/ and /a/. The location of this boundary correspond to the auditory value a_b for which the two corresponding auditory characterizations are equal, in other words, a_b satisfies:

$$P([A_\Phi = a_b] \mid [\Phi = /\varepsilon/]) \quad = \quad P([A_\Phi = a_b] \mid [\Phi = /a/]). \tag{5}$$

Since auditory characterizations are assumed to be normal distributions, Eq (5) becomes:

$$\mathcal{N}(a_b; \mu_A^\varepsilon, \sigma_A^\varepsilon) = \mathcal{N}(a_b; \mu_A^a, \sigma_A^a), \tag{6}$$

$$\Rightarrow \quad \frac{1}{\sqrt{2\pi}\sigma_A^\epsilon} e^{-\frac{(a_b-\mu_A^\epsilon)^2}{2{\sigma_A^\epsilon}^2}} = \frac{1}{\sqrt{2\pi}\sigma_A^a} e^{-\frac{(a_b-\mu_A^a)^2}{2{\sigma_A^a}^2}}. \tag{7}$$

Taking the logarithm of Eq (7) allows to find the desired relation:

$$\log\sigma_A^\epsilon + \frac{(a_b-\mu_A^\epsilon)^2}{2{\sigma_A^\epsilon}^2} \quad = \quad \log\sigma_A^a + \frac{(a_b-\mu_A^a)^2}{2{\sigma_A^a}^2}, \tag{8}$$

$$\Rightarrow \mu_A^\epsilon \quad = \quad a_b \pm \sigma_A^\epsilon \sqrt{2\log\frac{\sigma_A^a}{\sigma_A^\epsilon} + \frac{(a_b-\mu_A^a)^2}{{\sigma_A^a}^2}}. \tag{9}$$

Since the boundary we are interested in is located between μ_A^ϵ and μ_A^a, among the two solutions given by Eq (9), only the one for which $\mu_A^\epsilon < a_b$ is correct. Therefore:

$$\boxed{\mu_A^\epsilon = a_b - \sigma_A^\epsilon \sqrt{2 \log \frac{\sigma_A^a}{\sigma_A^\epsilon} + \frac{(a_b - \mu_A^a)^2}{{\sigma_A^a}^2}}.} \tag{10}$$

The value of a_b is obtained from the position of the auditory characterizations in normal conditions. Since auditory characterizations have the same variance in the normal condition, a_b is simply located at half the distance between the centers of the /ɛ-a/ region in normal condition. In other words:

$$a_b = \frac{{\mu_A^\mathcal{E}}^{(n)} + {\mu_A^a}^{(n)}}{2} = \frac{2+3}{2} = \frac{5}{2}. \tag{11}$$

Of course, the relation of Eq (10) does not need to be interpreted as a precise, necessary condition for our results to hold. We do not propose, either, that it is a property of the cognitive system of phoneme representation and learning. In other words, we do not claim that Eq (10) has to be satisfied or is a property of how phonemes representations are updated. Instead, we just computed the relation between the updated parameters $\mu_A^\mathcal{E}$ and $\sigma_A^\mathcal{E}$ such that the boundary between phonemes /a/ and /ɛ/ do not move at all. It is quite likely that repeated exposure to a shifted /ɛ/ will both decrease variance (as new samples have no variability) and displace the mean (towards the new location). These two mechanisms could exactly follow Eq (10), in which case the boundary would exactly not move. But, it is much more likely that they do not perfectly cancel out; in the experimental data, the boundary position probably moved, maybe under the measurement precision, and, in all likelihood, it moved less than in other experimental conditions.

Supporting information S5.
From L-14 to S-09: variations around the theme.

An apparent contradiction remains between the experimental data of S-09 and L-14. In L-14, the perceptual boundary shift resulting from motor learning occurs in the auditory region related to the adapted utterances of the subjects. In S-09 it occurs in the auditory region corresponding to what subjects hear during adaptation. This discrepancy was pointed out in L-14 but it was suggested that a possible explanation could lie in differences between sibilants, used in S-09, and vowels, used in L-14.

We have shown that our model is able to account for observations in L-14. Slight differences in the way the auditory-motor internal model and/or the auditory characterization of vowel /ɛ/ are updated could enable the prediction of observations in S-09. This is illustrated by the results presented in Fig 1, in the context of our three retained hypotheses.

On the one hand, if motor learning induces only a local update of the auditory-motor internal model in the context of a speech perception process involving the fusion of sensory pathways (Hypothesis $Q^{F}_{\text{Per}} \oplus H^{M}_{\text{Ad}}$), a slight displacement of the motor region affected by this local update in the direction of the auditory perturbation (Fig 1, top panel) would predict observations in S-09. This could be due to the fact that the extent of the articulatory changes due to compensation for the auditory perturbation is less important than in L-14. Such a difference in values of the model parameters predicting observations in L-14 is consistent with the fact that in S-09 the articulation of the fricative /s/ is at the front boundary of the articulatory space, which intrinsically limits the magnitude of the articulatory changes in the front direction, while the vowel /ɛ/ used in L-14 is articulated in the center of the articulatory space.

On the other hand, if motor learning induces both motor and auditory updates in the context of a pure auditory speech perception process (Hypothesis $Q^{A}_{\text{Per}} \oplus H^{M\Phi}_{\text{Ad}}$), shifting the auditory characterization of the perturbed phoneme combined with an insufficient narrowing (so that the opposite effects on the boundary shift would not cancel each other) would predict observations reported in S-09 (Fig 1, middle horizontal panel). Finally, combining both differences in the updates of the auditory-motor internal model and the auditory characterization of the perturbed phoneme would also predict observations in S-09 under hypothesis $Q^{F}_{\text{Per}} \oplus H^{M\Phi}_{\text{Ad}}$ (Fig 1, lower panel). Hence, in the context of our model, both observations in S-09 and L-14 could be explained by the same influences of motor learning on speech perception. Their differences would not be contradictory, they would only show that the two tasks induced differences in the amplitudes of the updates associated with motor learning.

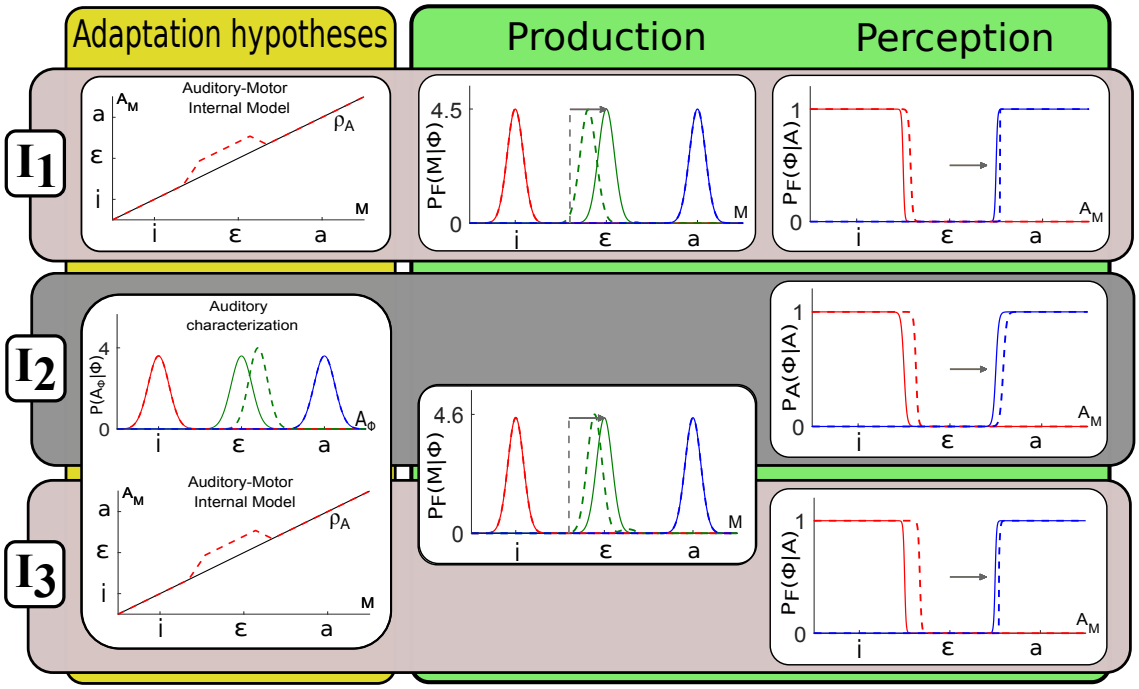

Figure 1: Proposed values of model parameters selected to account for perceptual boundary shifts on both sides of vowel /ɛ/ along the /i-a/ continuum. Each of the three mechanisms proposed above in the paper is able to account for such boundary shifts. For $Q^{F}_{\text{Per}} \oplus H^{M}_{\text{Ad}}$, if the locality of update of the internal model extends more to the /ɛ-a/ continuum compared to our previous simulations, then the perceptual shift also extents to this portion of the space. In $Q^{A}_{\text{Per}} \oplus H^{M\Phi}_{\text{Ad}}$ and $Q^{F}_{\text{Per}} \oplus H^{M\Phi}_{\text{Ad}}$, a perceptual boundary shift is obtained in both sides of vowel /ɛ/ along the auditory continuum with same shift but smaller narrowing than in previous simulations.

Titre — Modélisation Bayésienne de la planification motrice de la parole : variabilité, buts multisensoriels et intéractions perceptuo-motrices

Résumé — Dans cette thèse nous étudions certains aspects de la variabilité dans la production de la parole, avec comme point de départ la variabilité observée chez un locuteur dans la répétition d'un même son dans les mêmes conditions, que nous appelons variabilité intrinsèque. Les modèles de contrôle moteur de la parole abordent principalement la variabilité contextuelle de la parole mais prennent rarement en compte sa variabilité intrinsèque, alors même que l'on sait que c'est cette variabilité qui donne à la parole tout son caractère naturel. Dans le contexte général du contrôle moteur, l'origine précise de la variabilité intrinsèque reste peu comprise et controversée. Cependant, une hypothèse courante est que la variabilité intrinsèque serait essentiellement due à du bruit neuronal dans la chaine d'exécution.

L'objectif principal de cette thèse est d'aborder la variabilité intrinsèque et contextuelle de la production de la parole dans un cadre formel intégrateur. Pour cela nous faisons l'hypothèse que la variabilité intrinsèque n'est pas que le résultat d'un bruit d'exécution, mais qu'elle résulte aussi d'une stratégie de contrôle où la variabilité inter-répétition fait partie intégrante de la représentation de la tâche.

Nous formalisons cette idée dans un cadre computationnel probabiliste, la modélisation Bayésienne, où l'abondance de réalisations possibles d'un même item de parole est représentée naturellement sous la forme d'incertitudes, et où la variabilité est donc manipulée formellement. Nous illustrons la pertinence de cette approche à travers trois contributions.

Dans un premier temps, nous reformulons un modèle existant de contrôle optimal de la parole, le modèle GEPPETO, dans le formalisme probabiliste et démontrons que le modèle Bayésien contient GEPPETO comme un cas particulier. En particulier, nous illustrons comment l'approche Bayésienne permet de rendre compte de la variabilité intrinsèque tout en incluant les mêmes principes d'émergence et de structuration de la variabilité contextuelle proposés par GEPPETO.

Dans un deuxième temps, le formalisme nous permet de dépasser le cadre de GEPPETO en y intégrant une composante somatosensorielle dans la représentation des buts. Cela permet d'introduire une variabilité interindividuelle sur la préférence sensorielle, c'est-à-dire la modulation des poids relatifs des cibles auditives et somatosensorielles, et permet d'expliquer la variabilité de compensation observée dans les études de perturbation sensorielle. Cette étape a nécessité l'élaboration d'hypothèses sur l'intégration des retours sensoriels dans la planification, dont nous avons cherché à évaluer la pertinence en concevant une expérience originale de production-perception de parole.

Dans un troisième temps, nous exploitons le formalisme pour réinterpréter des données expérimentales récentes qui mettent en évidence un changement perceptif consécutif à un apprentissage moteur induit par une altération du retour auditif. Cela est rendu possible grâce à la représentation unifiée des connaissances dans le modèle, qui permet d'intégrer la production et la perception dans un cadre formel unique.

L'ensemble de ces travaux illustre la capacité du formalisme Bayésien à proposer une démarche systématique et structurée pour la construction des modèles. Cette démarche facilite le développement des modèles et leur complexification progressive en précisant et explicitant les hypothèses formulées.

Mots clés — Modèle computationnel, Modélisation Bayésienne, Parole, Contrôle moteur, Variabilité, Buts multisensoriels, intéractions pérceptumotrices.

Title — Bayesian modeling of speech motor planning: variability, multisensory goals and perceptuo-motor interactions

Abstract — In this thesis we study certain aspects of speech variability, our starting point being the variability characterizing the repetitions of a given utterance by a given subject, in a given condition, which we call intrinsic variability. Models of speech motor control have mainly focused on the contextual aspects of speech variability, and have rarely considered its intrinsic component, even though it is this fundamental component of variability that gives speech it naturalness. In the general context of motor control, the precise origin of the intrinsic variability of our movements remains controversial and poorly understood, however, a common assumption is that intrinsic variability would mainly originate from neural and muscular noise in the execution chain. The main goal of this thesis is to address the contextual and intrinsic component of speech variability in an integrative computational framework . To this aim, we postulate that the main component of the intrinsic variability of speech is not just execution noise, but that it results from a control strategy where intrinsic variability characterizes the abundance of possible productions of the intended speech item.

We formalize this idea in a probabilistic computational framework, Bayesian modeling, where the abundance of possible realizations of a given speech item is naturally represented as uncertainty, and where variability is thus formally manipulated. We illustrate the pertinence of this approach with three main contributions.

Firstly, we reformulate in Bayesian terms an existing model of speech motor control, the GEPPETO model, and demonstrate that this Bayesian reformulation, which we call B-GEPPETO, contains GEPPETO as a particular case . In particular, we illustrate how the Bayesian approach enables to account for the intrinsic component of speech variability while including the same principles proposed by GEPPETO for the emergence and structuration of its contextual component. Secondly, the Bayesian framework enable us to go beyond and extend B-GEPPETO in order to include a multisensory characterization of speech motor goals, with auditory and somatosensory components. We apply this extension to explore variability in the context of compensations to sensory-motor perturbation in speech production. We account for differences in compensation as sensory preferences implemented by modulating the relative contribution of each sensory modality in the model. The somatosensory characterization of speech motor goals involved a certain number of hypotheses that we intended to evaluate with two experimental studies. Finally, in our third contribution we exploit the formalism for the reinterpretation of recent experimental observations concerning perceptual changes following speech motor adaptation to auditory perturbations. This original analysis is made possible thanks to the unified representation of knowledge in the model, which enables to account for production and perception processes in a single computational framework.

Taken together, these contributions illustrate how the Bayesian framework offers a structured and systematic approach for the construction of models in cognitive sciences . The framework facilitates the development of models and their progressive complexification by specifying and clarifying underlying assumptions.

Keywords — Computational modeling, Bayesian modeling, Speech, Motor control, Variability, Multi-sensory goals, Perceptuo-motor interactions.