COSMO : un modèle bayésien des fondements sensorimoteurs de la perception et de la production de la parole

Jean-Luc Schwartz, Marie-Lou Barnaud, Pierre Bessière, Marc-Antoine Georges, Raphaël Laurent, Clément Moulin-Frier, Mamady Nabé, Jean-François Patri, Pascal Perrier, Julien Diard

COSMO : un modèle bayésien des fondements sensorimoteurs de la perception et de la production de la parole

Jean-Luc Schwartz[1], Marie-Lou Barnaud, Pierre Bessière[2], Marc-Antoine Georges[1,3], Raphaël Laurent, Clément Moulin-Frier[4], Mamady Nabé[1,3]
Jean-François Patri, Pascal Perrier[1], Julien Diard[3]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab (2) Sorbonne Univ., CNRS, ISIR
(3) Univ. Grenoble Alpes, CNRS, LPNC (4) Flowers Bordeaux Ensta ParisTech INRIA
jean-luc.schwartz@grenoble-inp.fr, julien.diard@univ-grenoble-alpes.fr

RÉSUMÉ

Nous avons développé un cadre de modélisation des processus de la communication parlée, COSMO (« Communicating Objects by Sensory-Motor Operations »), qui s'applique à la fois aux modèles de perception et de production de parole en installant les relations sensori-motrices au cœur de son programme. COSMO permet de formaliser conjointement deux cadres théoriques majeurs des recherches sur la communication parlée, les théories auditives et motrices – mais aussi de les intégrer au sein de théories perceptuo-motrices. Ceci conduit ainsi à de nouveaux modèles de perception alliant traitements auditifs et prise en compte de connaissances motrices, ou de nouveaux modèles de contrôle moteur de la parole orientés vers la réalisation de buts sensoriels multimodaux. Nous présentons ces avancées ainsi que des pistes de développement sur le traitement temporel et l'implémentation *deep learning* permettant d'aller vers l'apprentissage sur des données réelles.

ABSTRACT

COSMO: a Bayesian model of the sensory-motor bases of speech perception and production
COSMO ("Communicating Objects by Sensory-Motor Operations") is a framework for jointly modeling speech perception and production by considering sensory-motor relations as a core component of its program. COSMO allows to jointly formalize two major theoretical frameworks in speech research that are auditory and motor theories – but also to integrate them within perceptual-motor theories. This leads to new perception models associating auditory processing and motor knowledge, and new speech motor control models oriented toward the achievement of multimodal sensory goals. We present the main results obtained with COSMO, and perspectives about temporal processing and deep learning implementation allowing to get closer to learning on real data.

MOTS-CLÉS : Perception, production, théories auditives, motrices, perceptuo-motrices
KEYWORDS: Perception, production, auditory theories, motor theories, perceptual-motor theories

1 Introduction

La question de la nature des unités phonologiques et des représentations auditives, articulatoires ou motrices, mises en jeu dans les processus de perception et de production de la parole, a longtemps buté sur le manque de modèles computationnels permettant de formaliser et tester les hypothèses dans

un cadre comparatif, pour dépasser le stade des raisonnements théoriques et se confronter aux données expérimentales de manière quantitative. Nous avons développé depuis plus de 10 ans un tel cadre de modélisation, COSMO, qui fournit une architecture sensori-motrice adaptable à la fois aux questions portant sur la perception et sur la production de la parole. COSMO a permis de proposer des réponses concrètes à ces questions, avec des avancées qui nous semblent fortes, et conduisent globalement à une vision sensori-motrice cohérente des processus de la communication parlée.

2 COSMO, une architecture sensori-motrice pour la parole

COSMO repose sur une conception simple d'un acte langagier, résumée par l'acronyme « *Communicating Objects by Sensory-Motor Operations* », selon laquelle un locuteur transmet une information, dénommée de manière générale un « objet de communication », en produisant, par une action motrice M, un son S. L'action M réfère à l'objet O_M pour le locuteur, le son S réfère à l'objet O_S pour l'auditeur. La communication est supposée réussie lorsque les deux objets O_M et O_S sont identiques, et cette réussite est représentée par une variable booléenne C. L'hypothèse de base de COSMO est que chaque agent de cette interaction, locuteur comme auditeur, internalise intégralement cet échange (Fig. 1, gauche). Ainsi, dans un acte de communication, chaque interlocuteur est un « agent COSMO » équipé d'un modèle complet de la communication reposant sur les 5 variables du modèle (C, O_M, S, M, O_S), reprenant, au passage l'acronyme COSMO. Le cadre d'implémentation et d'utilisation de ce modèle s'inscrit dans le contexte global de la modélisation probabiliste bayésienne développé par Bessière et al. (2013). Ainsi, le modèle COSMO est décrit par une distribution de probabilité $P(C, O_M, S, M, O_S)$. Une série d'hypothèses d'indépendance conditionnelle conduit (Moulin-Frier et al., 2012 ; Laurent et al., 2017) à la structure de base d'un agent COSMO :

$$P(C, O_M, S, M, O_S) = P(O_M)\, P(M \mid O_M)\, P(S \mid M)\, P(O_S \mid S)\, P(C \mid O_M, O_S) \qquad (Eq.\ 1)$$

Dans cette équation, $P(M \mid O_M)$ représente les liens entre objets et actions motrices du système moteur, $P(S \mid M)$ est le modèle interne « direct » permettant à l'agent d'associer les sorties sonores S correspondant aux commandes motrices M, $P(O_S \mid S)$ est le décodeur sensoriel accédant aux objets à partir des sons et $P(C \mid O_M, O_S)$ est le modèle de communication donnant à la variable booléenne C une valeur 1 si les objets O_M et O_S sont identiques (Fig. 1, droite).

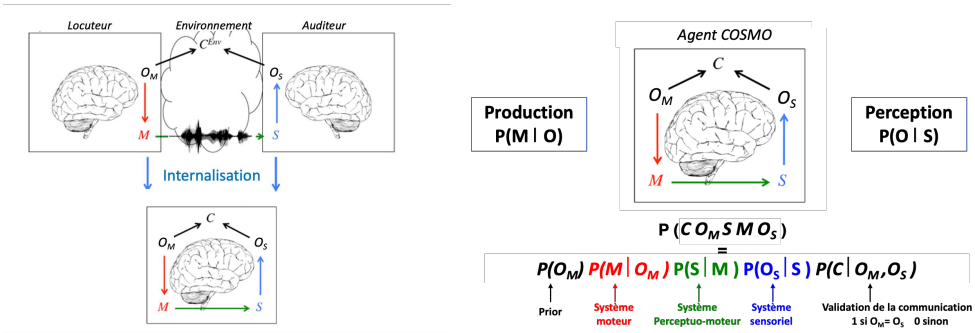


FIGURE 1 : Le modèle COSMO

L'« agent COSMO » défini par l'Eq. (1) peut, dans le cadre du raisonnement bayésien (Bessière et al., 2013), répondre à des questions probabilistes caractérisant son comportement dans des tâches cognitives données. Ainsi, une tâche de production est définie par la question probabiliste $P(M \mid O)$: « sachant un objet O, quel geste moteur M lui associer ? ». Et une tâche de perception est définie par la question probabiliste $P(O \mid S)$: « étant donnée une entrée sensorielle S, quel objet O lui associer ? ».

C'est ainsi que la distribution de probabilité de l'Eq. (1) permet de résoudre à la fois des tâches de perception et de production, ce qui correspond à la première originalité mentionnée en introduction. Mais un autre intérêt majeur de l'hypothèse d'internalisation de COSMO et de sa formalisation par l'Eq. (1) est que chacune de ces tâches peut être conçue au moins de deux manières différentes, selon que l'objet O ci-dessus renvoie à l'auditeur (O_S) ou au locuteur (O_M). Et précisément, ce choix réfère à deux cadres théoriques majeurs qui ont animé les débats sur la nature des représentations des unités de la parole dans la perception et la production depuis plus de 60 ans, comme on le voit sur la Fig. 2.

Cette figure déroule le raisonnement bayésien dans des tâches de production et de perception dans le contexte des théories motrices ($O=O_M$), auditives ($O=O_S$) mais aussi des théories perceptuo-motrices qui apparaissent comme opérant une simple fusion probabiliste des mécanismes de ces deux théories. Ainsi, dans le contexte des théories motrices, la perception de la parole implique une inversion sensori-motrice permettant d'inférer les gestes à partir des sons, associée à un décodage moteur (comme dans Liberman & Mattingly, 1985) ; et la production de la parole est organisée directement autour d'objets moteurs sans référer à la valeur auditive des gestes produits, comme dans la phonologie articulatoire de Browman & Goldstein (1989). Dans le contexte des théories auditives, la perception de la parole implique un accès direct aux relations entre signaux et sens sans référer à quelque connaissance motrice que ce soit (Diehl et al., 2004) et la production de la parole implique le contrôle par un modèle direct permettant d'assurer la réalisation de cibles auditives, comme dans les premières versions du modèle DIVA (Guenther et al., 1998). Enfin, les théories perceptuo-motrices de la perception de la parole considèrent à la fois des processus de catégorisation auditive directe et de complément d'information par simulation motrice (comme dans la Perception-for-Action Control Theory, PACT, de Schwartz et al., 2012) ; et les processus de contrôle associent accès aux répertoires moteurs et spécification sensorielle des cibles (voir Guenther et al., 2006 ; ou Perrier et al., 2005).

3 COSMO appliqué à la perception de la parole

3.1 « *Auditory-Narrow Motor-Wide* » : le rôle de l'inférence motrice dans le bruit

Les données de neuroimagerie montrent que les régions corticales motrices (dans le cortex frontal) jouent un rôle spécifique dans le traitement perceptif de la parole dans le bruit ou le traitement de stimuli atypiques. Elles sont plus activées dans ces conditions et semblent effectivement participer de manière importante au processus d'analyse phonétique et de décodage (Du et al., 2014). Or il n'y a pas réellement, à notre connaissance, d'explication convaincante à cette capacité spécifique du système moteur à traiter de stimuli atypiques ou bruités. Nous avons étudié cette question avec COSMO. Pour ce faire nous avons exposé un agent COSMO tel que défini sur la Fig. 1 à un processus d'apprentissage dans lequel il apprend les trois distributions de l'Eq. (1), répertoire moteur $P(M \mid O_M)$, modèle interne « direct » $P(S \mid M)$, décodeur sensoriel $P(O_S \mid S)$, d'une manière semi-supervisée qui semble cognitivement plausible. Dans ce processus semi-supervisé, l'agent apprenant reçoit de son environnement des paires (S, O_S) fournies par un maître (parent par exemple). L'hypothèse sous-jacente est que l'environnement peut fournir à l'agent en développement des données sensorielles S et des informations complémentaires permettant de communiquer de manière synchrone l'objet $O_M = O_S$ ainsi désigné. Il doit alors inférer la variable non fournie : le geste moteur correspondant M. Ce processus d'inférence (Laurent et al., 2017) permet de mettre à jour les trois distributions initialement non connues. L'agent peut ensuite utiliser ces distributions pour activer, en phase de perception de parole, un processus d'inférence auditive $P(O_S \mid S)$ et un processus d'inférence motrice $P(O_M \mid S) = \Sigma_M (P(M \mid O_M)\ P(S \mid M)$, puis, dans le cadre d'une théorie perceptuo-motrice comme celle que nous défendons (Schwartz et al., 2012), fusionner les résultats de ces deux inférences (voir Fig. 2).

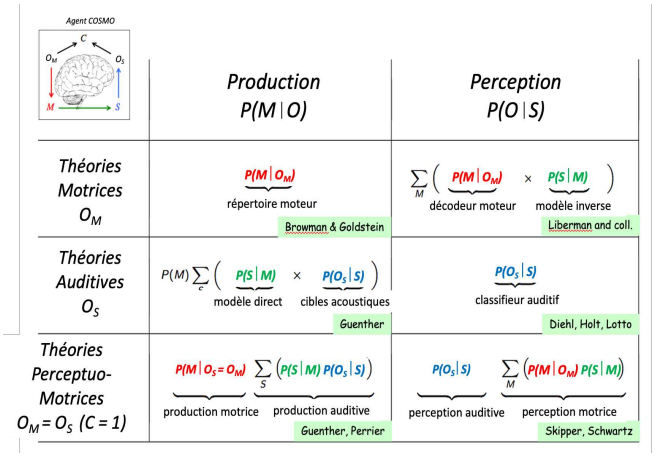


FIGURE 2 : La formalisation COSMO des modèles de production et perception dans le cadre des théories motrices, auditives et perceptuo-motrices (d'après Laurent et al., 2017)

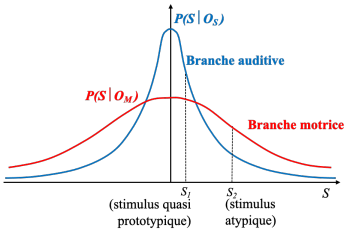


FIGURE 3 : La propriété « Auditory-Narrow Motor-Wide »

Or les simulations montrent que, si l'apprentissage de la distribution du décodeur sensoriel $P(O_S \mid S)$, entièrement supervisé, est rapide et bien focalisé autour des stimuli d'apprentissage, les distributions impliquant la variable M du modèle interne $P(S \mid M)$ et du répertoire moteur $P(M \mid O_M)$ sont apprises de manière non supervisée, plus lente et moins efficace, ce qui rend la branche de décodage moteur moins précise pour traiter des stimuli prototypiques. Par contre, de manière frappante, cette branche a été confrontée au cours de son apprentissage à des stimuli auto-générés par l'agent puisque non fournis par l'environnement, et ces stimuli, souvent non prototypiques, permettent au décodeur moteur d'être plus robuste à des entrées non standard et notamment dans le bruit. On peut résumer cette propriété en comparant les distributions $P(S \mid O_S)$ et $P(S \mid O_M)$ qui reflètent les représentations perceptives respectivement produites par la branche auditive et la branche motrice de décodage, et constater que la représentation auditive est plus précise que la représentation motrice, donc rend mieux compte des stimuli prototypiques (comme S1, Fig. 3) mais moins bien des stimuli atypiques (comme S2, Fig. 3). C'est ce que nous avons désigné sous le terme de propriété « *Auditory-Narrow Motor-Wide* », ANMW (Laurent et al, 2017). Nous avons récemment pu confirmer cette hypothèse par des données de neuroimagerie IRMf dans un paradigme de répétition-suppression : si l'on présente à des participants des trains de 4 voyelles identiques, la réponse corticale (signal BOLD) est plus faible que si la dernière occurrence est différente des trois premières (par des phénomènes de fatigue/adaptation nerveuse). Mais la modulation qu'il faut appliquer sur le 4ème stimulus de la séquence doit être plus grande pour produire une différence significative dans les aires motrices frontales que dans les aires auditives temporales, qui apparaissent ainsi plus sélectives (Dole et al., soumis). Ainsi la propriété computationnelle ANMW semble correspondre à une propriété fonctionnelle du cortex humain.

3.2 Caractérisation auditive et motrice des sons de parole

Plutôt que d'opposer la recherche d'invariants articulatoires/moteurs et auditifs pour caractériser les unités de la parole, COSMO a permis de montrer une complémentarité entre les deux voies (Laurent et al., 2017) : si les voyelles sont finalement bien caractérisées pour leurs propriétés acoustiques et notamment leurs formants, les plosives, à cause de phénomènes de coarticulation bien connus, posent un problème majeur pour la recherche d'invariants associés à leur lieu d'articulation. Par contre, si un agent COSMO a accès à la fois aux propriétés auditives (S) et motrices (M) associées à une classe

phonétique donnée, alors il peut associer à la variabilité des formes acoustiques d'une plosive en contexte comme /b/ ou /d/ des propriétés articulatoires contrastives (geste labial vs. dental) qui permettent de leur associer effectivement un marqueur invariant. Nous avons montré que des bébés de 9 mois, lorsqu'ils commencent à babiller, semblent capables d'associer des gestes et des sons et présentent un premier accès à une propriété sensori-motrice invariante associée à un lieu d'articulation donné (/b/ vs. /d/). Cette capacité n'était pas présente sur des bébés plus jeunes (6 mois) ou non babillants, ni sur des sons non disponibles dans les premiers temps du babillage (Vilain et al., 2019).

3.3 Autres études en perception de parole

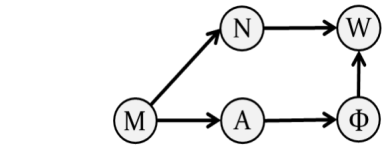
Nous avons mené d'autres travaux sur les propriétés des branches auditive et motrice du décodage perceptif et leur intégration dans une théorie perceptuo-motrice comme celle de la PACT. Nous avons pu ainsi montrer (Barnaud et al., 2019) que l'existence d'une branche motrice de décodage permettait d'expliquer l'existence d'idiosyncrasies couplées en perception et en production, comme celles observées par Ménard & Schwartz (2014) sur l'organisation perceptive des voyelles. Chaque locuteur en français présente une organisation spécifique de ses niveaux de hauteur vocalique, certains locuteurs avec une proximité acoustique des voyelles hautes et mi-hautes (/i/ et /e/), d'autres des voyelles mi-hautes et mi-basses (/e/ et /ɛ/) ou mi-basses et basses (/ɛ/ et /a/). Or cette spécificité en production se retrouve en miroir dans les processus de caractérisation perceptive par ces mêmes sujets. Le passage par une branche de décodage moteur $P(S \mid O_M)$ rend compte de cette propriété. Nous avons montré globalement la cohérence de COSMO avec un ensemble de données neurocognitives et proposé une architecture corticale plausible pour le modèle (Barnaud et al., 2018).

4 COSMO-production : le modèle *Bayesian GEPPETO*

4.1 Principes de base de *Bayesian GEPPETO*

Le cadre COSMO a également inspiré des développements en production de parole, en lien avec le modèle GEPPETO (« ***GE****stures shaped by the* ***P****hysics and by a* ***PE****rceptually oriented* ***T****argets* ***O****ptimization* ») organisant le contrôle autour de cibles perceptives (Perrier et al., 2005). GEPPETO actionne les déplacements de la langue dans le conduit vocal par des commandes musculaires simulées par un modèle à éléments finis et suivant les principes de la théorie du point d'équilibre (Perrier et al., 2005). Dans GEPPETO, une action est ainsi définie à la fois par une cible auditive associée à un objectif phonologique (un phonème ou une succession de phonèmes) et par une modélisation de la dynamique du système périphérique de production, modulée par un niveau de force musculaire variable (associé à des contraintes de stress, de prosodie et de clarté d'articulation).

Une reformulation probabiliste de GEPPETO a conduit Patri et al. (2015) à proposer un nouveau modèle, *Bayesian GEPPETO*, à l'architecture proche de celle de COSMO. Une première brique d'architecture est présentée sur la Fig. 4 (gauche) : un geste moteur M est caractérisé à la fois par un objectif auditif A associé à un phonème Φ et par la spécification du niveau de force N associé à un niveau d'effort discrétisé W (de type « faible, moyen, fort »). La distribution de probabilité caractérisant cette brique de base est indiquée sur la figure. Cette distribution permet de piloter le modèle vers des cibles auditives, conformément au cadre théorique. L'enchaînement de cibles Φ^1, Φ^2, Φ^3 dans le temps (Fig. 4, droite) génère des phénomènes de coarticulation par le jeu de contraintes d'économie articulatoire spécifiées dans le lien entre les commandes motrices successives M^i par une « variable de cohérence » booléenne C_M en relation avec l'effort souhaité W^i (Patri et al., 2015).



$$P(M\ A\ \Phi\ N\ W) = P(M)\ P(A \mid M)\ P(\Phi \mid A)\ P(N \mid M)\ P(W \mid N\ \Phi)$$

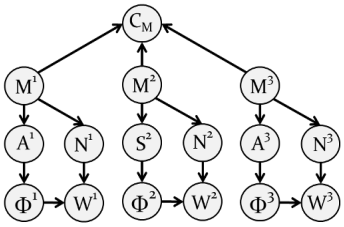
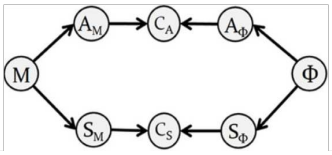


FIGURE 4 : La première version du modèle Bayesian *GEPPETO* (d'après Patri, 2018).

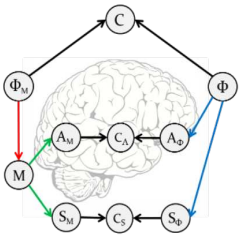
4.2 Un contrôle moteur gérant la relation entre cibles auditives et articulatoires

Comme pour la perception de la parole, les modèles de production de parole peuvent associer des objectifs auditifs et articulatoires dans les processus de contrôle moteur. Les analyses de la littérature (Patri et al., 2019) montrent en effet clairement qu'il existe à la fois des données en faveur des cibles auditives (comme celles des « lip-tube » de Savariaux et al. (2005) où le blocage de la cible articulatoire conduit à une réorganisation visant des objectifs auditifs) ; mais également des mises en évidence de l'existence d'objectifs articulatoires dans la production, comme dans l'étude de Tremblay et al. (2003) où une perturbation dynamique de la mâchoire sans conséquences acoustiques est prise en compte et produit des modifications du contrôle. Les principaux modèles de production de la parole prennent en compte ces deux séries d'objectifs, auditifs mais aussi somatosensoriels – la somesthésie permettant de traduire les objectifs articulatoires en entrées sensorielles. C'est ainsi le cas de DIVA (Guenther et al., 2006) comme de *Bayesian GEPPETO* (Patri et al., 2019).

L'existence de cette double voie de contrôle permet de définir dans *Bayesian GEPPETO* des commandes motrices *M* qui prennent en compte à la fois des objectifs auditifs *A* et somatosensoriels *S*, chacun reliés à l'objectif phonétique Φ (Fig. 5, gauche). Ce cadre a permis à Patri et al. (2019) de rendre compte de différences entre locuteurs, « plutôt auditifs » ou « plutôt somatosensoriels » dans leurs réponses à des perturbations (Lametti et al., 2012). Il a permis également à Patri et al. (2018) de fournir un système d'explications cohérentes à un ensemble de données expérimentales sur l'adaptation audio-motrice, dans lesquelles il a été montré que l'apprentissage moteur en parole peut altérer les frontières catégorielles entre les sons. Et, finalement, Patri (2018) montre comment, par une expansion de *Bayesian GEPPETO* intégrant un lien direct entre commandes motrices et spécification phonétique (Fig. 5, centre), on peut aller vers une spécification à la fois auditive (*A*), articulatoire/somatosensorielle (*S*) et motrice (*M*) des unités phonétiques (Φ), avec des poids variables pour chacune des modalités selon les conditions de production et la fiabilité accordée à chaque canal sensoriel, en cohérence avec le cadre perceptuo-moteur de COSMO (Fig. 5, droite).



$$P(M\ \Phi A_M A_\Phi\ C_A\ S_M\ S_\Phi\ C_S) = P(M)\,P(\Phi)\,P(A_M \mid M)\,P(A_\Phi \mid \Phi)\,P(C_A \mid A_M\ A_\Phi)\,P(S_M \mid M)\,P(S_\Phi \mid \Phi)\,P(C_S \mid S_M\ S_\Phi)\ .$$



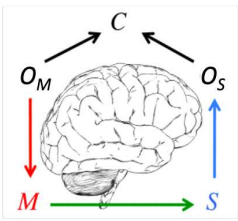


FIGURE 5 : *Bayesian GEPPETO* dans sa version auditive-somatosensorielle (gauche, d'après Patri et al., 2019) et le lien avec le cadre général de COSMO (à droite, d'après Patri, 2018).

4.3 Autres études en production de parole

Patri et al. (2020) ont montré, par un paradigme original, qu'on peut conduire un participant à positionner sa langue pour donner à son conduit vocal la forme adoptée lors de la production d'une voyelle sans qu'il ait eu d'information sur cette voyelle, puis juger sa faculté à identifier la voyelle sur la seule base des retours somatosensoriels. Les données confirment, de manière très convaincante, une telle capacité de « catégorisation somatosensorielle ». Nous avons fait également une excursion dans le domaine de la morphogenèse des systèmes sonores des langues du monde et montré comment faire émerger, dans une société d'agents sensori-moteurs COSMO, des systèmes vocaliques et consonantiques cohérents avec les données disponibles (Moulin-Frier et al., 2015).

5 Perspectives

5.1 *COSMO-Onset*, l'interfaçage avec les rythmes neuronaux

Les neurosciences cognitives progressent à grand pas dans la compréhension des processus d'analyse et de contrôle des réseaux corticaux. Dans le domaine de la perception de la parole, l'architecture rythmique/hiérarchique de Giraud & Poeppel (2012) est devenue une pièce centrale de tout travail de modélisation, proposant notamment que les rythmes theta proto-syllabiques (typiquement entre 4 et 8 Hz) structurent temporellement le traitement cortical de l'information auditive. Nous avons entrepris d'inclure cette hypothèse de contrôle temporel au sein du modèle *COSMO-Onset* (Nabé et al., 2021). Ce modèle (Fig. 6) combine un module de contrôle temporel et un module de décodage, plus classique. Le module de contrôle temporel combine par une variable de cohérence C un système de détection des cycles proto-syllabiques, basé sur l'analyse bottom-up de l'enveloppe acoustique, avec les prédictions top-down sur les structures rythmiques de la langue. Sur cette base, il ouvre et ferme des « portes bayésiennes » qui gèrent le flux d'information dans le module de décodage. Ce système, en développement, permet potentiellement de rendre compte de données comme celles d'Aubanel & Schwartz (2020) montrant que le traitement de la parole dans le bruit bénéficie à la fois de propriétés rythmiques (avantage aux séquences isochrones) et de connaissances linguistiques (avantage aux rythmes naturels par rapport à des rythmes modifiés).

5.2 *Deep-COSMO*, le passage à la réalité des données via le *deep learning*

Toutes les simulations menées sur les différentes versions de COSMO ont porté jusqu'à présent sur des données de parole de synthèse, de structures simples, qui ont permis de mettre au point les concepts et de poser les bases des raisonnements. Mais il est essentiel de passer à l'échelle pour se confronter à des situations réelles et notamment pour évaluer les mécanismes d'apprentissage sur de la parole naturelle. Pour ce faire, nous développons avec T. Hueber et L. Girin, spécialistes d'apprentissage machine, des architectures neuronales sensori-motrices permettant d'apprendre des représentations par un processus d'apprentissage intégral. Nous présentons sur la Fig. 7 le dernier avatar de ces développements. Dans cette architecture associant variables sensorielles et articulatoires (Georges et al., 2022), nous avons d'abord configuré un système articulatoire neuronal réaliste (Georges et al., 2020) capable de générer du son $\hat{s}$ à partir des configurations articulatoires a (fonction $\boldsymbol{\Phi}$ sur le modèle). Le système peut alors apprendre conjointement, à partir de données acoustiques s (typiquement des paramètres cepstraux) un modèle interne direct $\boldsymbol{f}$, associant les commandes articulatoires a à des sorties estimées $\hat{s}$, et un modèle inverse $\boldsymbol{g}$ permettant d'inférer les commandes a à partir des entrées s. L'apprentissage est réalisé par une séquence de passes d'apprentissage conjointes portant alternativement sur le modèle direct et sur le modèle inverse. L'apprentissage du

modèle direct minimise l'erreur (« *loss* ») entre le son produit $\tilde{s}$ et le son estimé $\hat{s} = \boldsymbol{f}(a)$ pour une configuration articulatoire a, afin d'apprendre la fonction $\boldsymbol{f}$. L'apprentissage du modèle inverse minimise l'erreur entre objectif acoustique s et signal inféré $\hat{s} = \boldsymbol{f}(a)$ en gelant le modèle direct (la fonction $\boldsymbol{f}$) pour centrer la rétro-propagation du gradient sur l'apprentissage de la fonction $\boldsymbol{g}$. Cette stratégie d'apprentissage dite *end-to-end* fournit des performances d'apprentissage conjoint non supervisé de modèle direct et inverse convaincantes. Le modèle est ainsi capable d'imiter la parole de plusieurs locuteurs, avec des performances globales satisfaisantes, validées à la fois par une évaluation objective (décodage par HMM) et subjective (tests d'écoute) (Georges et al., 2022).

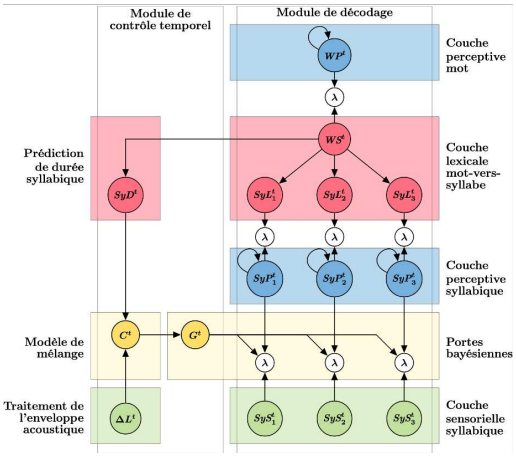


FIG. 6 : *COSMO-Onset* (Nabé et al., 2021)

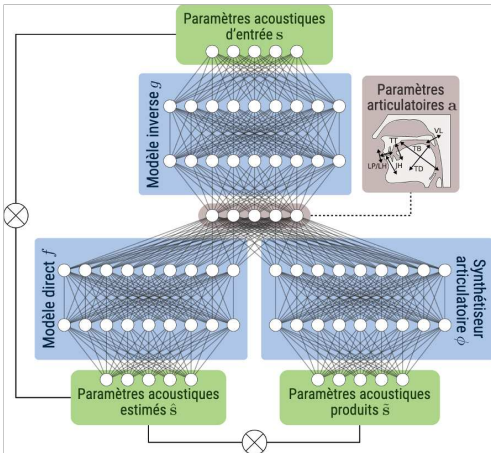


FIG. 7 : *Deep-COSMO* (Georges et al., 2022)

Remerciements

Ce travail a bénéficié du soutien du Multidisciplinary Institute in Artificial Intelligence MIAI@Grenoble-Alpes (ANR-19-P3IA-0003) et du projet ISP Université Grenoble Alpes Bio-Bayes Predictions (ANR-15-IDEX-02; Auvergne-Rhône-Alpes (AURA) Region PAI-19-008112-01 grant).

Références

AUBANEL V., SCHWARTZ J.L. (2020). The role of isochrony in speech perception in noise. *Sci Reports*, 19580.

BARNAUD M.L., BESSIÈRE P., DIARD J., SCHWARTZ J.L. (2018). Reanalyzing neurocognitive data on the role of the motor system in speech perception within COSMO, a Bayesian perceptuo-motor model of speech communication. *Brain & Language* 187, 19-32,

BARNAUD M.L., DIARD J., BESSIÈRE P., SCHWARTZ J.L. (2019). Computer simulations of coupled idiosyncrasies in speech perception and speech production with COSMO. *PloS ONE* 14, e0210302.

BESSIÈRE P., MAZER E., AHUACTZIN J.M., MEKHNACHA K. (2013). *Bayesian programming*. Boca Raton, Florida: CRC Press.

BROWMAN C.P., GOLDSTEIN L. (1989). Articulatory gestures as phonological units. *Phonology* 6, 201–251

DIEHL R., LOTTO A., HOLT L. (2004). Speech perception. *Annual Review of Psychology* 55, 149–179.

DOLE M., ET AL. (SOUMIS). Comparing the selectivity of vowel representations in cortical auditory vs. motor areas: A repetition-suppression study.

Du Y., Buchsbaum B.R., Grady C.L., Alain C. (2014). Sensorimotor integration aids speech perception. *PNAS* 111, 7126-7131.

Georges M.A., Badin P., Diard J., Girin L., Schwartz J.L., Hueber T. (2020). Towards an articulatory-driven neural vocoder for speech synthesis. Actes de *ISSP 2020*.

Georges M.A., Diard J., Girin L., Schwartz J.L., Hueber T. (2022). Repeat after me: self-supervised learning of acoustic-to-articulatory mapping by vocal imitation. Actes *ICASSP 2022* (à paraître).

Giraud A.L., Poeppel D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nat. Neurosci* 15:511.

Guenther F.H., Ghosh S.S., Tourville J.A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang* 96, 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, 611–633.

Lametti, D. R., Nasir, S. M., Ostry, D. J. (2012). Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. *J. Neuroscience* 32, 9351–9358.

Laurent R., et al (2017). The complementary roles of auditory and motor information evaluated in a Bayesian perceptuo-motor model of speech perception. *Psychological Review* 124, 572-602

Liberman A.M., Mattingly I.G. (1985). The motor theory of speech perception revised. Cognition 21, 1–36.

Ménard L., Schwartz J.L. (2014). Perceptuo-motor biases in the perceptual organization of the height feature in French vowels. *Acta Acustica* 100, 676-689.

Moulin-Frier C., et al. (2012). Adverse conditions improve distinguishability of auditory, motor and perceptuo-motor theories of speech perception. *Lang. and Cognitive Processes* 27, 1240–1263.

Moulin-Frier C., Diard J., Schwartz J.L., Bessière P. (2015). COSMO: a Bayesian modeling framework for studying speech communication and the emergence of phonological systems. *Journal of Phonetics* 53, 5–41.

Nabé, M., Schwartz, J.L., Diard, J. (2021). COSMO-Onset: A Neurally-Inspired Computational Model of Spoken Word Recognition, Combining Top-Down Prediction and Bottom-Up Detection of Syllabic Onsets. *Frontiers in Systems Neuroscience* 15, pp.653975.

Patri J.F. (2018). *Bayesian modeling of speech motor planning: variability, multisensory goals and perceptuo-motor interactions*. Thèse, Université Grenoble Alpes.

Patri J.F., Diard J., Perrier P. (2015). Optimal speech motor control and token-to-token variability: a Bayesian modeling approach. *Biological Cybernetics* 109, 611.

Patri J.F., Diard J., Perrier P. (2019). Modeling sensory preference in speech motor planning: a Bayesian modeling framework. *Frontiers in Psychology* 10, 2339.

Patri J.F., et al. (2020). Speakers are able to categorize vowels based on tongue somatosensation. *PNAS* 117, 6255-6263.

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 Comp. Biol. 14, e1005942.

Perrier P., Ma L., Payan Y. (2005). Modeling the production of VCV sequences via the inversion of a biomechanical model of the tongue. Actes Interspeech 2005 (Lisbon), 1041—1044.

Savariaux, C., Perrier, P., Orliaguet, J.-P. (1995). Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: a study of the control space in speech production. *JASA* 98, 2428–2442.

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, 336-354.

Tremblay S., Shiller D.M., Ostry D. J. (2003). Somatosensory basis of speech production. *Nature* 423, 866–869.

Vilain, A., et al. (2019). The role of production abilities in the perception of consonant category in infants. *Developmental Science* 22, pp.e12830.