Mémoire en vue d'une
Habilitation à Diriger des Recherches

Représentations phonétiques et technologies vocales

Gérard BAILLY
Institut de la Communication Parlée
UPRESA CNRS n°5009
INPG/ENSERG - Université Stendhal

septembre 99

Remerciements

La rédaction d'un document aussi personnel permet de mettre en lumière des traits de ma perception de la recherche scientifique très contradictoires au premier abord : la nécessité d'une démarche personnelle, curieuse, ouverte et exigeante mais aussi celle d'un projet collectif, cohérent et synergique. C'est l'idée de la nécessaire adéquation entre ambition personnelle et aventure collective. C'est l'idée d'un même objet éclairé par de multiples sources, d'un modèle façonné par de multiples mains... qu'il serait souvent bien difficile de démêler, sur les sillons gravés dans l'objet, les mains, les voix qui ont mis en forme ces lignes.

Ce mémoire suit certes une ligne directrice qui motive la presque totalité de mon travail sur la parole : la mise en valeur voire l'émergence de niveaux de représentations phonétiques. Ces "grilles" de lecture motrices, acoustiques voire perceptives de la parole me semblent être le bon premier niveau d'interface entre la matière brute, le monde sensible et nos représentations mentales, le monde construit. L'émergence de ces représentations depuis la matière brute n'a rien à attendre d'une naïve spontanéité du réel : c'est la nécessaire conception de champs d'expériences qui donneront aux constructions échafaudées sur ces représentations un fondement scientifique.

Ce goût immodéré pour les représentations phonétiques est peut-être le fruit d'une rencontre entre un cursus d'ingénieur et un objet de recherches accessible par la mesure, mais plus sûrement celui de rencontres avec des chercheurs, ou plus simplement des hommes ou des femmes que je veux remercier ici :

René Carré, pour m'avoir détourné du CERN, m'avoir convaincu de me présenter au CNRS et pour avoir créé l'ICP dans sa forme,

Jean-Marc Dolmazon, pour avoir créé ses institutions, permis son épanouissement et l'émancipation de ses chercheurs et symbolisé sa réussite,

Pierre Escudier, pour savoir écouter le cœur et le cerveau de ce laboratoire, pour vouloir les défendre tous deux et savoir diffuser l'information,

Christian Abry, dont le cœur est palpable, l'esprit agile et l'œil vif, pour ouvrir des espaces vers lesquels je mets hélas tant de temps à pousser le regard,

Véronique Aubergé qui tente, et réussit souvent, à rendre ma boulimie de travail inutile, pour son goût immodéré des métaphores et de l'optimalité,

Pierre Badin et Louis-Jean Boé, qui ont toujours le même âge, pour leur amour immodéré des livres et des données,

Frédéric Berthommier, le bouillon de culture et le collègue d'à-côté, pour sa disponibilité, son ouverture et sa curiosité toujours en éveil,

Rafaël Laboissière, sincère, ambitieux et brillant, qui pousse des chercheurs moins ambitieux mais aussi moins brillants à se remettre en question,

Nino Medves, l'homme en bleu, pour son dévouement, la constance de son soutien logistique et de ces compétences techniques,

Pascal Perrier, qui ose poser les questions dont on ne connaît pas la réponse, pour son audace scientifique et la qualité de son encadrement,

Jean-Luc Schwartz, qui m'a fait enfin comprendre le sens d'une démonstration par l'absurde, et qui sait faire donner le meilleur d'eux-mêmes aux pires défaitistes,

Tous les autres collègues de l'ICP, joueurs de foot ou non, qui rament chacun à leur rythme et à leur style, de manière à ce que l'aventure continue…

Pierre-François Marteau, Thierry Barbe, Haïdong Wang, Mhania Guerti, Mamoun Alissali, Plinío Barbosa, Adrien Neagu, Yann Morlec et Bleike Holm, dont le travail commun et les thèses jalonnent ce mémoire,

Béatrice, Juliette et Maxime pour être tout à tour le refuge, le port d'attache, le téléphone impatient, le rire absolu et la gorge serrée, pour savoir dormir plus que moi et me réveiller à temps. Sans oublier Hélène, René et Jean-Pierre qui ne me quittent jamais,

Enfin, je dois remercier Jacqueline Vaissière, Mario Rossi, Daniel Hirst, Mary Beckman, Nick Campbell et tant d'autres chercheurs rencontrés dans les projets et dans les congrès internationaux qui ont influencé ma recherche, parfois par leur position antagoniste sur des objets partagés, souvent par les ouvertures qu'ils m'ont suggérés...

Ces remerciements ont été les premiers mots de ce mémoire que j'ai rédigé péniblement en quelques trois ans d'inscriptions renouvelés à l'INPG. A cette époque, Christian Benoît était encore parmi nous. Il était le véritable chef d'une équipe soi-disant bicéphale. Par son enthousiasme, la cohérence de son travail et son humour, il était une formidable machine à penser, à entreprendre et à créer. Je réalise maintenant toutes les pistes de recherches qu'il avait su ouvrir… J'ai perdu aussi un diable de complice et je n'oublierai jamais notre visite à Woodstock.

Table des matières

Table des illustrations

« *En un mot, j'ai mis mon orgueil à prouver, ici, mon ignorance, aux admirables savants qui honorent notre espèce.* »
[villiers:93] *p.228.*

Avant-propos

J'ai été longtemps réfractaire à l'idée de devoir passer de nombreuses heures à sécher devant un écran blanc pour accoucher de ces quelques lignes. Le regard sur le passé imposé par cet exercice m'a semblé être une perte de temps qui me semblait plus profitablement investi dans cet encadrement dont on vise ici la reconnaissance. Ce changement de position est essentiellement motivé par deux raisons : d'une part, la maturité du projet scientifique de l'Institut de la Communication Parlée doit être accompagnée par une responsabilisation de ses acteurs ; ces mêmes acteurs doivent se donner les moyens de défendre et de promouvoir ce projet collectivement et individuellement. D'autre part, après avoir travaillé avec les jeunes chercheurs dont le travail commun a servi de support à ce manuscrit, et vu les affres des échéances finales, je me devais d'essayer de mettre en cohérence ces travaux et, étant a fortiori moins forcé par le temps et les financements, me soumettre au même travail et au même jugement.

« *Ainsi, j'eusse blâmé, par exemple, le Phonographe de son impuissance à reproduire, en tant que bruits, le bruit… de la chute de l'empire romain… les bruits qui courent…les silences éloquents… et, en fait de voix, de ce qu'il ne peut clicher ni la voix de la conscience?… ni la voix -- du sang?… ni tous ces mots merveilleux qu'on prête aux grands hommes… ni le Chant du cygne… ni les sous-entendus… ni la voix lactée? Non! Ah! Je vais trop loin.* »
[villiers:93] *p.45.*

INTRODUCTION

Le monde sensible de la parole

La parole est et restera (il semble que le téléphone cellulaire couvrira la planète d'une maille tellement fine que monde réel et monde sensible ne feront plus bientôt qu'un[1]) le moyen le plus efficace pour transmettre le langage. Au delà de ce code très conventionnel bien qu'en perpétuelle évolution, la parole permet aussi un « double » langage : la prosodie, ou manière de « parler », nous permet de nous situer par rapport au discours, et donc, par cela même, parfois d'en nier la véracité. Elle est aussi et peut-être avant tout un signal biologique, où une partie de notre corps et souvent notre âme résonnent...

Pour communiquer avec autrui, notre corps créée, structure et émet donc une multitude de signaux que nous avons la capacité d'enregistrer, stocker et analyser de plus en plus massivement et de plus en plus finement : cet éventail couvre les signaux neurophysiologiques, leurs conséquences motrices analysées en termes de déplacements, vitesses... de structures osseuses, de cartilages, de tissus musculaires, de graisse (!) ou de peau, ou bien encore leurs conséquences aérodynamique et acoustique. L'ensemble de ces signaux constitue donc l'ensemble du monde sensible de la parole accessible à l'expérience.

Organiser le monde sensible de la parole

L'ensemble de ces signaux est donc porteur de sens, véhicule un flux d'informations de natures diverses en direction ou non d'un interlocuteur. L'établissement, la gestion de la communication langagière passe donc par un processus d'encodage, de transmission et de décodage d'unités d'informations qui permettront aux interlocuteurs de modifier et d'enrichir leur connaissance, leurs "représentations mentales" de l'autre, des autres ou de leur environnement. Cette mise en relation de représentations mentales suppose que les interlocuteurs partagent au moins deux types de représentations : une représentation mentale - ou plutôt un ensemble de conventions de représentation - permettant à l'information incidente de signifier quelque chose, et une représentation « phonologique » permettant d'associer à cette représentation une suite de sons distinctifs. On retrouve ici les niveaux d'articulation de Martinet, monèmes et phonèmes, qui jalonnent la création et la mise en forme d'un message. A l'instar de l'impossibilité d'une association directe des monèmes à un ensemble de signaux sonores, peut-on considérer un rapport direct entre la représentation phonologique du langage et signaux sensibles?

[1] Tout voir, tout sentir, tout goûter, tout entendre sans être là…

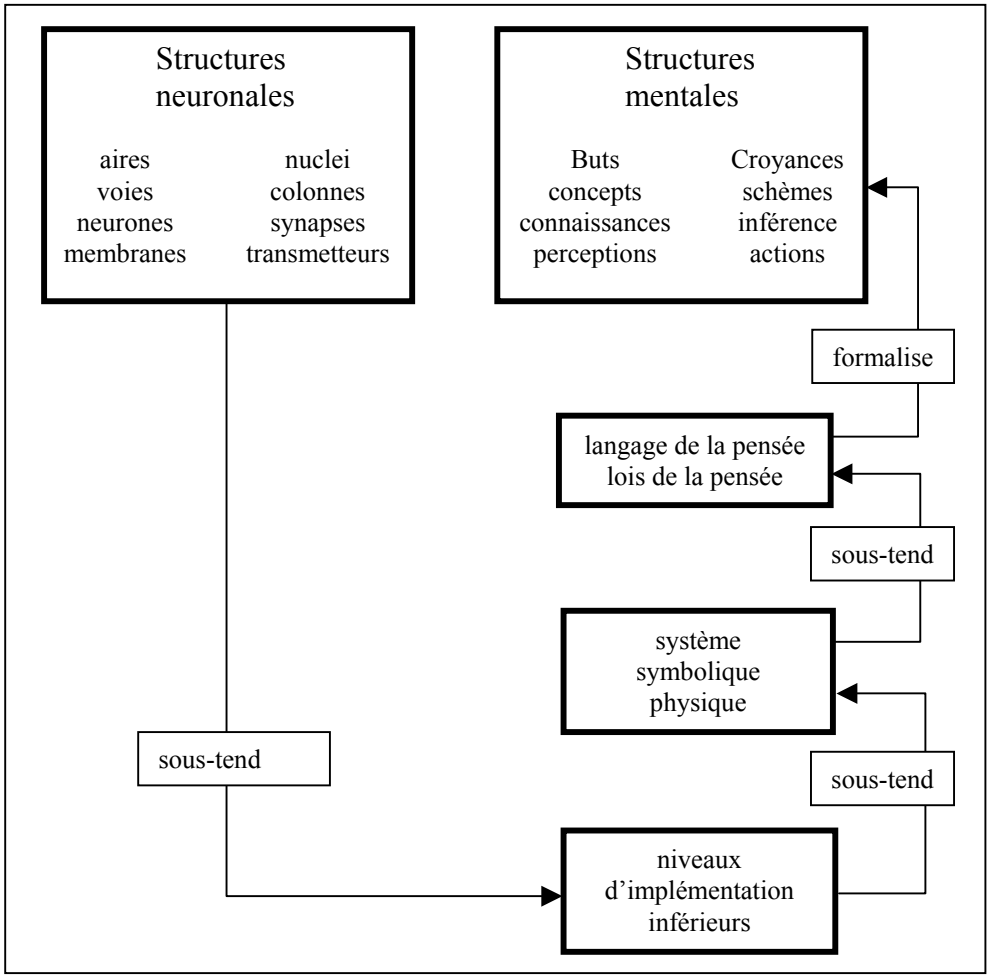

Figure 1: Structures neronales et mentales dans le paradigme symbolique (d'après [smolensky92](p.80)).

Les technologies vocales : la tentation sous-symbolique

Les technologiques vocales ont connu ce que Hofstadter appelle » le rêve booléen » : de nombreux systèmes de reconnaissance et de synthèse ont cherché à décrire par des règles les relations complexes entre des structures mentales et les signaux physiques. Comme le souligne Smolensky (voir Figure 1), cette approche sous-tend l'existence d'un système symbolique physique procédant par calcul sur des unités sémantiquement interprétables (systèmes de traits, d'indices...). Il est difficile d'approcher ainsi le monde physique et d'extraire des signaux de grande variabilité des descripteurs suffisamment robustes et fins pour que l'interprétation sémantique puisse être convenablement initiée : les systèmes sont rigides et fragiles. Le paradigme sous-symbolique (voir Figure 2) introduit par Smolensky vise à distribuer l'information et à remplacer la notion d'unités sémantique par la notion de formes ou de configurations d'activité et d'états dynamiques. Le symbole n'est alors qu'un étiquetage, une formalisation abstraite d'un processus dynamique qui exploite la notion d'état sans avoir besoin de leur associer une sémantique propre ni leur associer un niveau d'interprétation. On retrouve ici la notion d'état dans une chaîne de Markov cachée ou d'unité cachée dans un réseau de neurone.

Cette approche connaît un grand succès dans les technologies vocales et « dévore » peu à peu les domaines de recherche jusqu'alors traités par l'approche symbolique. Cette tendance est illustrée par le développement des méthodes globales en reconnaissance et synthèse de parole jusque dans les modèles de langage ou de dialogue.

Dans les modèles classiques de reconnaissance par Chaînes de Markov Cachées, le signal acoustique est décomposé de manière aveugle en une séquence de trames de paramètres spectraux souvent enrichis de leurs premières ou secondes dérivées, chaque phonème étant conçu comme une suite d'états dont la seule différentiation est leur faculté d'émettre de manière privilégiée certains types de trames. Si les réseaux de neurones ont la faculté de "calculer" dans les couches dites cachées des représentations du signal intermédiaires entre la caractérisation du signal d'entrée - évidemment, elle aussi codant de manière uniforme la composition spectrale du signal- et la caractérisation symbolique de sortie, il est difficile d'interpréter ces représentations sinon de manière triviale. La force de ces approches globales est de proposer une capacité d'apprentissage automatique permettant à des structures de représentation très peu différenciées de s'adapter de manière implicite et surtout optimale - au regard de critères objectifs - aux signaux à représenter.

Le développement des méthodes globales en synthèse est plus récent : bien que les méthodes de synthèse dites par concaténation d'unités aient été développées dès les années 60, la nécessité d'un contrôle, d'un traitement prosodique de la parole synthétique relègue encore l'association entre la chaîne phonétique et un ensemble de signaux pré-stockés à une simple production de matière sonore brute. Le développement récent de systèmes d'étiquetage prosodique et la possibilité de traiter automatiquement et de stocker d'importantes quantités de signaux ont contribué à faire émerger une réponse technologique à la question précédente [campbell:atr96] : la synthèse de parole peut être vue comme un accès linguistiquement contrôlé à de gigantesques dictionnaires de sons. Le problème de modélisation du signal et de manipulation de ses caractéristiques structurelles fait ici place à la caractérisation sémantique de ce dernier ou plus précisément à son étiquetage au regard de sa fonction au sein du système de représentations phonologiques, voire même directement cognitives, sous-jacentes. L'avantage d'une telle stratégie de génération de signaux est de ne nécessiter pratiquement aucune intervention sur les signaux de parole originaux pré-stockés, ce qui permet de préserver de manière implicite la naturalité, et de ne rompre qu'occasionnellement et de manière très locale la cohérence des signaux synthétiques : en effet, les techniques de modification des caractéristiques des sons ne garantissent pas que les signaux résultants puissent effectivement avoir été produits par le locuteur d'origine. Le processus de synthèse des signaux se réduit alors à un simple compromis entre adéquation à la tâche et problèmes de concaténation.

Le paradigme morphologique

De telles approches se heurtent d'un simple point de vue technologique à des limites évidentes de généricité, interpolation et extrapolation depuis des données brutes qui souffrent et souffriront toujours d'incomplétude et de manque de signification statistique au vu du nombre de paramètres à estimer (voir [bailly:atr97]).

D'où l'idée que ce problème d'apprentissage mal posé pouvait être mieux appréhendé par des contraintes posées a priori. D'où peuvent venir ces contraintes? Si elles ne peuvent pas être apprises à partir du sens, elles doivent provenir du système assurant le transport du sens, le système d'encodage et de décodage de l'information. Les signaux sont en effet soumis aux exigences des systèmes de production et de perception de parole, qui sont soumis aux lois "universelles" de la gravité, de la propagation des ondes acoustiques... et aux limites des systèmes biologiques en termes de fréquence de coupure et de temps de réponse que ce soit en émission ou en réception. Ainsi les systèmes dynamiques ou les formes d'activité des modèles

sous-symboliques ne s'organisent pas de manière quelconque sur une tabula rasa. L'apprentissage est heureusement contraint : nous disposons de guides.

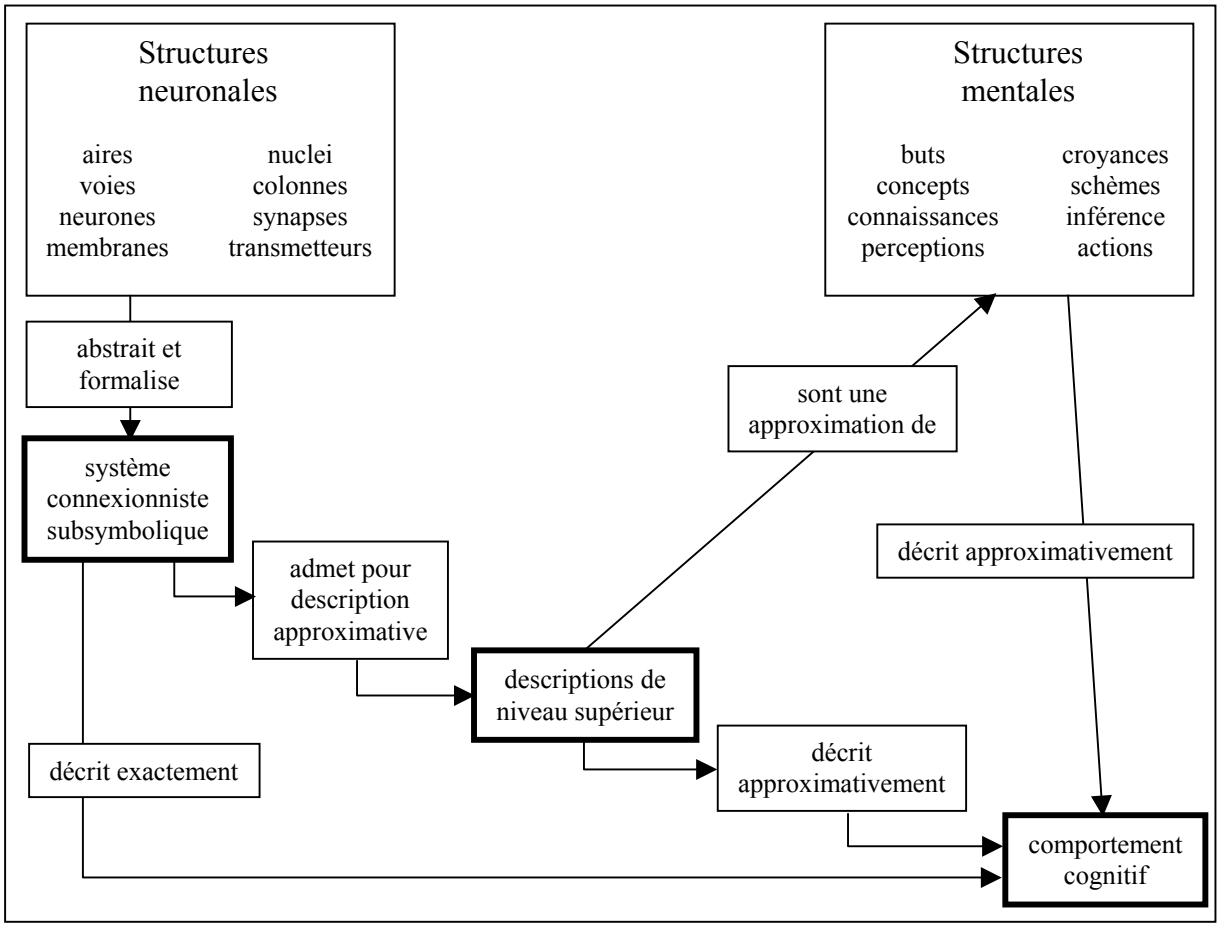

Figure 2: Structures neuronales et mentales dans le paradigme sous-symbolique (d'après [smolensky92]p.83).

L'influence des lois et des contraintes physiques régissant le monde physique sur notre entendement, sur la perception de ce même monde physique est une problématique qui beaucoup plus généralement intéressé de nombreux philosophes comme Kant ou Husserl, logiciens comme Wittgenstein, linguistes comme Jackendoff et mathématiciens comme Thom. Un des élèves de René Thom, Jean Petitot, invité par Jean-Luc Schwartz, fit un séminaire à Grenoble en 1987 dont le contenu et les références ont inspiré, souvent de manière sournoise, une grande partie de mes travaux et ont constitué une ligne directrice que je vais tenter de clarifier ici. Je ne le rends évidemment pas responsable de la mésinterprétation de son travail mais un schéma de principe tracé à cette occasion a éclairé le mien : au centre de son travail [petitot:85] [petitot:86] [petitot:rs90] est la réinterprétation de la phénoménologie de Husserl très projective et de la conception gestaltiste à la lumière de la Théorie des Catastrophes de Thom. Selon Husserl, un objet doit "se détacher en tant que phénomène" [Husserl; 1972; p.26]. Ainsi, les objets que nous percevons dans le monde sensible ne sont pas uniquement le fruit d'une décision arbitraire purement systémique mais doivent posséder des qualités d'*immédiateté* perceptive des "Gestaltqualitäten" de von Ehrenfels. Dans son schéma synoptique des divers niveaux de passage du monde sensible au monde mental, Jean Petitot introduit pour ce faire un niveau dit *morphologique*, réceptacle des qualités intrinsèques du signal acoustique et des attentes construites par nos représentations mentales. Ces attentes sont rendues possibles grâce à la morphogenèse du sens : le signifié est construit à partir de structures catastrophiques préexistantes liées aux

diverses transformations neurales, motrices, proprioceptives et perceptives que les signaux de contrôle de la parole subissent. La constitution d'un code phonologique procèderait alors par « phagocytage », par amplification d'une fonction, de phénomènes somatiques plus fondamentaux à l'instar de l'hypothèse des marqueurs somatiques d'A. Damasio [damasio:95] : la fonction de régulation biologique, d'identification d'émotions primaires du système limbique est « capturée », réutilisée par le système associatif du cortex frontal afin d'associer à un événement des émotions dites secondaires. Ce mécanisme permet d'associer à une contingence d'images corticales correspondant à un événement une qualité bonne ou mauvaise en regard de la survie, tant au niveau de la bio-régulation que de l'interaction sociale de manière à conduire efficacement une prise de décision dans ces domaines.

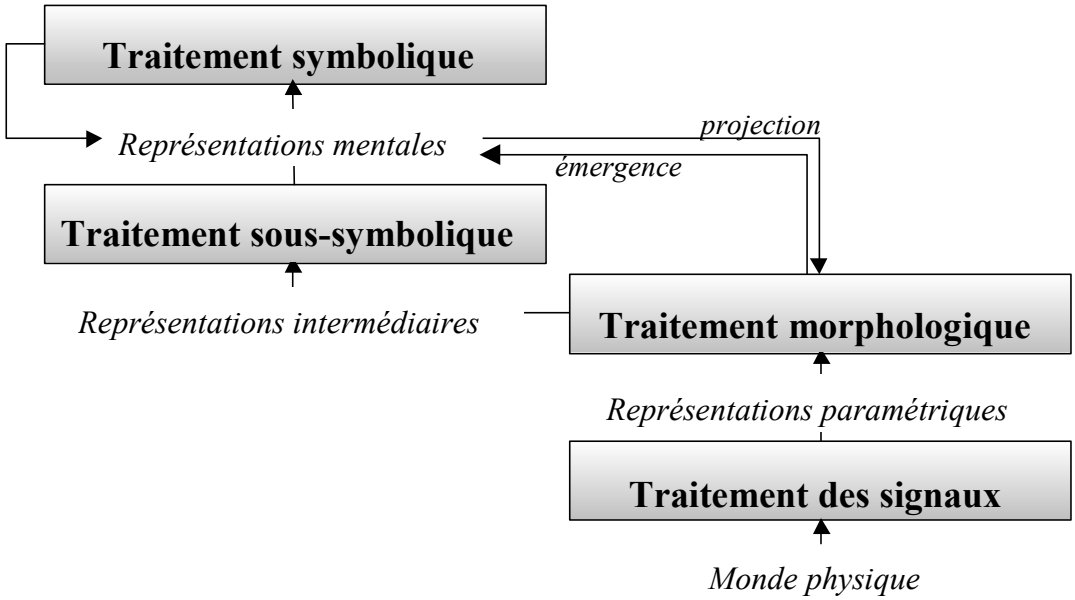

Figure 3: le traitement morphologique dans le processus d'observation du monde physique. Les représentations mentales émergent en prenant appui sur nos capacités naturelles à produire et à percevoir des différences. La catégorisation émerge alors d'une auto-organisation contrainte par ces frontières naturelles préexistantes et du nécessaire accroissement de l'entropie (nb. symboles à transmettre avec max. de fiabilité). En retour ces représentations mentales projtettent et ainsi façonnent de nouvelles frontières.

Bien sûr, il semble difficile a priori de séparer les contributions du substrat neural, moteur ou perceptif et des contraintes systémiques dans la morphogenèse du sens. Certaines théories et résultats expérimentaux viennent cependant étayer la rentabilité de cette approche, où les objets phonologiques viennent exploiter et enrichir des structures "catastrophiques" préexistantes et où l'observation phénoménologique par la statistique ne permet de remonter souvent qu'à des structures « catastrophiques » fossiles.

La théorie quantique de Stevens et ses prolongements

Pour Husserl, la saillance phénoménale est directement liée à des discontinuités physiques :

> « C'est à partir d'une limite de l'espace ou du temps que l'on saute d'une qualité à une autre. Dans ce passage continu d'une partie d'espace à une autre partie d'espace, nous ne progressons pas d'une manière également continue dans la qualité qui les recouvre, mais (...) à un endroit de l'espace où les qualités limitrophes ont un écart fini (et pas trop petit).» (Husserl 1972 p.29)

Ken Stevens propose en 1972 une illustration de ce propos. Stevens [stevens:72] [stevens:perilus91] propose ainsi que les objets phonétiques s'appuient sur une théorie de la

production de discontinuités spectrales et temporelles dans un continuum physique. Si son argumentaire est centré sur les non-linéarités du passage de l'articulatoire à l'acoustique, on peut aisément ajouter comme support à cette théorie les nombreux travaux sur les détecteurs de propriétés acoustiques ou perceptives. L'exemple le plus frappant est celui du F'2, mettant en exergue la faculté et/ou la propension du système auditif à extraire de manière cohérente et quantique des masses spectrales du signal acoustique. Cette contrainte perceptive a été exploitée avec succès dans la DFT, un système de prédiction des systèmes vocaliques d'inspiration morphogénétique [boe-etal:dft94]. Basé sur la théorie de la dispersion proposée par Liljencrants & Lindblom [liljencrants-lindblom:language72], il incorpore une contrainte additionnelle de focalisation : focalisation F1-F2 pour les voyelles arrières /a/ et /u/, F2-F3 (/y/) voire F3-F4 (/i/) pour les voyelles avant. Ce qui me semble être un indice fossile d'une ontogenèse morphologique est le fait que bien que l'opposition entre /i/ et /y/ en suédois et en français [schwartz-etal:jphon93] soit réalisée de manière différente, que les contraintes systémiques d'opposition soient différentes (présence en suédois du /u/ délabialisé) et qu'ainsi les cibles acoustiques soient différentes, les frontières perceptives soient exactement les mêmes et effectivement celle de la discontinuité maximale de la fonction F'2(F3).

Rythmicité et l'horloge interne

Les représentations de la structure rythmique dans les technologies vocales sont extrêmement pauvres : souvent limitée à des modèles corrélationnels entre durées segmentales. Ces modèles sont principalement exploités en synthèse via des associateurs linéaires [klatt:jasa87] [oshaughnessy:jphon81] [oshaughnessy:jasa84] ou linéaires par parties [bartkova-sorin:speechcom87] [riley:tm92] voire polynomiaux par parties [vansanten:tm92] [vansanten:csl94]. Dans tous les cas, un ensemble de prédicteurs phonotactiques, phonologiques, linguistiques voire paralinguistiques sont utilisés afin de prédire de manière optimale un ensemble de durées segmentales. De manière similaire en reconnaissance de parole, les modèles de phonèmes en contexte sont augmentés de modèles probabilistes de paramètres prosodiques associés, dont leur durée.

L'étude de la structure rythmique de la parole montre cependant que la syllabe (ou des évènements caractéristiques de la syllabe) jouent un rôle important dans l'organisation temporelle du signal de parole. Citons ici les travaux de Grosjean [gee-grosjean:cp83] [monnin-grosjean:ap93]sur les structures de performance et les travaux sur la synchronisation de la production de syllabes isolées [marcus:these76] [marcus:pp81] [morton-etal:pr76] [allen:jphon75] [fraisse:74] [fraisse:80] voire de séquences [berthier-etal:icphs91] avec d'autres activités motrices telles que taper ou synchroniser la production d'une syllabe avec d'autres signaux naturels ou synthétiques présentés périodiquement.

Le rôle prépondérant de l'oscillation mandibulaire sur la structuration et le développement du langage [davis-macneilage:ls94] [davis-macneilage:jshr95] n'est évidemment pas étranger à ce cadre d'analyse syllabique. L'exploitation de cette rythmicité « naturelle » est corroborée par les données sur l'acquisition du langage [konopczynski:icphs91][smith:jphon78] montrant que le babil infantile se démarque tardivement d'une isochronie initiale vers 16 mois pour incorporer la structure rythmique de la langue maternelle.

Certaines propositions vont encore plus loin et propose une régulation de la rythmicité syllabique par une horloge interne [turvey-etal:90] [hary-moore:biocyber87] . L'analyse statistique de grands corpus de parole fait d'ailleurs émerger des attracteurs syllabiques situés à des valeurs multiples d'une horloge de base d'environ 140~ms[fant-kruckenberg:kth89] [fant-etal:jphon91] [fant-kruckenberg:icslp96] soit 7~Hz valeur très proche de la fréquence propre de l'oscillation mandibulaire. Ces attracteurs « quantiques » sont aussi présents lorsque les pauses sont considérées. Dans son étude comparée de l'anglais, le français et le suédois, Fant note :

> « …the average inter-stress interval within a short time memory span of about 4 seconds preceding a pause… synchronises an internal beat generating clock which sets a preferred pause duration.» [p.248][fant:icphs91]

Cette tendance à reprendre la phonation à un instant en relation avec le tempo précédent - avec éventuellement des mesures « beats » silencieuses [fant-kruckenberg:icslp96] [barbosa-bailly:scom94] [barbosa-bailly:tm97] - a été aussi mis en évidence pour les tours de parole (« turn-taking ») par Couper-Kuhlen [couper-kuhlen:icphs91].

L'idée d'une médiation syllabique dans l'organisation temporelle compte cependant de nombreux détracteurs et de multiples réfutations d'une trop stricte régularité du tempo ont été opposées tant au niveau d'un cadre de programmation syllabique des durées segmentales, de la soit-disante classification des langues en regard de leur gestion de l'isochronie [dauer:jphon83] que de la synchronisation des tours de parole [bull:these97][2]. Si aucun corrélat physique d'une supposée régularité motrice ou perceptive n'a été mis à jour, il reste cependant à expliquer quel est le support de notre remarquable efficacité à prédire la suite d'une structure rythmique en parole comme en musique

Quelques pistes morphologiques en technologies vocales

La question reste cependant posée de savoir si ces propositions de structuration morphologique de la parole perçue par notre architecture motrice et les capacités de nos traitements perceptifs, sont rentables pour les technologies vocales.

Reconnaissance de parole et phonologie articulatoire

Les chaînes de Markov cachées sont des modèles de production : probabilités de transition et d'observation sont ajustées de manière à ce qu'un parcours de leurs états aient émis de manière optimale des suites d'observations d'apprentissage. La notion d'état peut être reliée à une morphologie du signal élémentaire où des zones temps/fréquence du signal acoustique et/ou visuel présentent une identité qualitative. La seule mesure de cette identité qualitative est donnée par une distribution optimale du signal sur un ensemble d'états. Cette distribution et la croissance éventuelle des états sont donc régis par les seules données. On peut cependant légitimement supposer que cette structure de surface du signal peut être prédite par une structuration articulatoire plus profonde. Comme nous le montrerons plus loin, cette structuration articulatoire peut certes se fixer comme but la facilitation du découpage acoustique du signal en contrôlant de manière plus ou moins précise le phasage des gestes sous-jacents. Cette structuration articulatoire permet cependant de déconvoluer les signaux en gestes organisés plus lents [mcgowan:scom94], plus homogènes et surtout plus génériques. La transformation articulatori-acoustique étant largement non-linéaire, un accès à des représentations articulatoires permet de réduire le maillage de l'espace de sortie et donc l'ensemble des exemplaires d'apprentissage : des phénomènes comme l'apparition de consones épenthétiques peuvent notamment être facilement modélisés comme des déphasages relâchés de gestes sous-jacents.

Au lieu d'envisager un ambitieux projet d'inversion, Li Deng et son équipe [erler-deng:csl93] [erler-freeman:jasa96] envisage l'introduction des connaissances phonétiques sur la gestuelle articulatoire en termes de structuration a priori des états : la phonologie articulatoire [browman-goldstein:85] [browman-goldstein:py86] [browman-goldstein:phono89] [browman-goldstein:jphon90] conçoit la programmation motrice en parole comme une suite de gestes intrinsèques (« core gestures ») chevauchants. Ces gestes spécifient l'extension spatio-temporelle d'un ensemble de variables de contrôle du conduit

[2] Les résultats obtenus par A. Capobianco en 1998 dans le cadre de son DEA confirment d'ailleurs cette absence de synchronisation dans les tours de parole.

vocal, assimilables à la spécification de constrictions. La sous-spécification de ces gabarits spatio-temporels permet de comprendre et gérer les anticipations motrices et moduler ces gabarits en fonction du style de parole. L'avantage de ce type de modèle est sa quasi-linéarité : connaissant des exemplaires aux extrêmes du comportement admissible (tout le problème reste quand même là!), on peut envisager d'interpoler des comportements tout à fait admissibles et ainsi obtenir un modèle de production robuste et générique. L'équipe de Li Deng démontre ainsi qu'une structure articulatoire générique peut être prédite pour un mot quelconque et convertie en un treillis d'états acoustiques homogènes directement interprétables par une chaîne de Markov cachées. La généricité du modèle permet alors de produire des modèles performants avec moins de corpus d'apprentissage. On peut même envisager de générer des données acoustiques absentes de ce corpus par synthèse… lorsque la synthèse de parole par modèles physiques aura atteint une généricité et une fidélité satisfaisantes.

Reconnaissance de parole et pré-traitements perceptifs

Kuhn [kuhn:jasa75] [kuhn:jasa79] puis Hermansky [hermansky-broad:icassp92] ont mis en valeur notre capacité à produire et percevoir des masses spectrales comme un des termes de la monnaie d'échange acoustique mis à disposition des systèmes phonologiques. La réduction de l'information spectrale nécessaire à la caractérisation de l'observation acoustique en reconnaissance de parole a bénéficié de ces recherches directement inspirées d'une approche morphologique : les deux concentrations spectrales suffisantes à caractériser une voyelle ont été modélisées par les 5 coefficients de l'analyse par prédiction linéaire sur la sortie d'un modèle d'oreille constituant l'analyse PLP. Complété dans l'analyse RASTA-PLP par l'introduction du masquage temporel [hermansky-etal:icassp95], ce type de paramétrage est à présent extrêmement populaire comme pré-traitement du signal acoustique.

De même, le paradigme CASA (pour « Computational Auditory Scene Analysis ») auquel la reconnaissance de parole va vraisemblablement emprunter de plus en plus de principes de traitement et de représentations, est directement inspiré de la vision et des principes de construction de représentation du monde sensible proposé par les Gestaltistes : extraction de primitives par des représentations spécialisées, perception de schémas…

Codage de parole et courbe de masquage

Les systèmes de codage de la parole utilisent un filtrage adaptatif de l'erreur de reconstruction du signal issu de résultats de psycho-acoustique voire de perception de parole : la représentation temps-fréquence initiale du signal de parole obtenu par les divers algorithmes de traitement du signal spécifiques ou non parole (FFT, LPC, cepstre, sinusoïdale, wignerville, ondelettes, formes d'ondes formantiques,…), souvent unique, exploitée par les systèmes de codage intègre souvent quelques propriétés tonotopiques de l'oreille interne (échelles non-linéaires de représentation des fréquences, courbes de masquage temporelles et fréquentielles…) qui permettent de mettre en forme les erreurs de modélisation de manière à les compenser de manière optimale par le système d'allocation de bits de quantification.

Il y a fort à parier que cette technologie devra de plus en plus intégrer les connaissances glanées sur les représentations perceptives (multimodales!) susceptibles de simuler notre capacité à analyser une scène auditive complexe où l'entropie n'est pas uniquement régie par l'énergie.

Synthèse de parole et analyse de la mélodie

Les traitements purement symboliques, la synthèse dite « par règles », ont largement dominé les efforts consacrés à la recherche en synthèse jusque dans les années 80

[klatt:jasa87] [carlson-granstrom:scom75] [carlson-granstrom:icassp76]. Le développement de compilateur de règles [carlson-granstrom:scom75] [hertz-etal:assp85] [vancoile:icassp89] auquel nous avons apporté notre contribution [alissali:these93] [bailly-alissali:ts92] en est la trace tangible (voir Annexe 1). Le système de règles avait alors en charge de « calculer » entièrement une représentation riche du signal de parole à partir d'une représentation phonologique de la tâche d'énonciation. La mise en cohérence temporelle et spatiale de ces divers paramètres est un problème complexe résolu de manière très diverse, soit par un enrichissement de la représentation phonologique et du mécanisme de projection de la structure phonologique sur la structure phonétique [local:94], soit par l'intégration de contraintes articulatoires dans le synthétiseur [stevens-bickley:jphon91].

Si on considère le simple adressage de signaux élémentaires, l'association d'un nœud de la structure phonologique à un exemplaire phonétique comme un traitement sous-symbolique élémentaire, la synthèse par concaténation résout la mise en cohérence temporelle et spatiale des signaux de manière triviale. Il reste que ces signaux peuvent être de structure plus riche qu'un simple enregistrement - par exemple, contenir des schémas plus complexes - et les règles et modalités d'enchaînement de ces exemplaires peuvent suivre des modèles plus élaborés qu'une simple concaténation. Citons dans cette veine, le modèle de synthèse développé par Coleman et Local [coleman:tm92] [local:94] où un schéma associant plusieurs paramètres acoustiques est associé à chaque allophone de l'arbre phonologique. Ces schémas sont alors combinés par chevauchement temporel et la trajectoire de chaque paramètre calculée de manière analogue à la phonologie articulatoire. Citons aussi le modèle de génération de l'intonation proposé par Aubergé [auberge:these91] où l'encodage des diverses fonctions linguistiques et paralinguistiques de l'intonation se réalise par une superposition de schémas prosodiques multiparamétriques.

Un programme de recherche centré sur les représentations morphologiques

Le paradigme de la représentation morphologique permet de travailler de manière claire sur la nature et la caractérisation de la monnaie d'échange utilisée par l'homme pour communiquer. La constitution de cette monnaie d'échange s'appuie sur les degrés de liberté anatomiques de nos organes de production, sur nos possibilités motrices, les capacités de nos organes perceptifs à traiter le monde physique et à structurer l'espace sensoriel résultant. Le pari de mon programme de recherche est de montrer que la structure du langage s'appuie sur ce cadre perceptuo-moteur, qu'elle exploite ou "pirate" des régularités ou des singularités des signaux biologiques.

L'outil essentiel de cette démonstration est la démarche expérimentale permettant de mettre en relation des données physiologiques, psycho-acoustiques et perceptives à des tâches linguistiques. La démonstration de l'intérêt de cette démarche peut être envisagée sous plusieurs formes :

(a) montrer la généricité des modèles élaborés : les modèles, après avoir été paramétrés sur des données, doivent pouvoir extrapoler des comportements et montrer ainsi que cet étage de représentation structure de manière cohérente les données brutes.

(b) montrer leur universalité : il se peut qu'un certain langage et qu'un locuteur dans une certaine situation n'exploite pas ou viole certaines contraintes perceptuo-motrices. On s'attend cependant que ces contraintes agissent comme des attracteurs ou qu'elles laissent des traces fossiles. Ces traces peuvent être d'ordre développemental comme le montre, par exemple, l'étude de l'apprentissage rythmique comparé entre enfant anglais et français [konopczynski:icphs91], les études de Davis & MacNeilage sur la structuration syllabique par l'oscillation mandibulaire [davis-macneilage:jshr95], ou les études de

« bootstraping » prosodique de l'apprentissage de la structure syntaxique [morgan-demuth:96].

(c) montrer une compréhension phénoménologique ou cognitive : nous avons utilisé et continuerons à utiliser des paradigmes de perturbation (de production ou de perception) pour montrer la non-transparence de ces niveaux de représentation morphologique. On voit bien comment l'effet Mc Gurk peut être exploité pour nous renseigner sur les modèles d'intégration audio-visuelle, comment la création de monstres acoustiques, comme l'a fait A. Neagu récemment [neagu-bailly:icslp98], peut suggérer des modèles de décodage des plosives ou comment les expériences de dévoilement progressif « gating » faites par Grépillat [auberge-etal:eurospeech97] peut nous renseigner sur les propriétés représentations phonétiques de l'intonation.

(b) montrer leur rentabilité technologique : une représentation doit pouvoir être évaluée dans des systèmes de synthèse, de codage ou de reconnaissance.

A travers deux thématiques de recherches, les chapitres suivants vont s'attacher à montrer que les recherches que j'ai menées et les travaux que j'ai encadrés peuvent s'organiser autour de ces quelques lignes.

« *Percevoir est tout autant une question d'action sur l'environnement que de réception des signaux en provenance de ce dernier… Il existe dans le cerveau des circuits neuraux qui élaborent en permanence une représentation de l'organisme, reflétant sa perturbation par des stimuli de l'environnement physique et socioculturel, et son action sur cet environnement. »*
[damasio:95] *p.284-285.*

STRUCTURATION DE L'ESPACE SENSORI-MOTEUR

Lors de mon intégration comme chargé de recherches au CNRS, l'ICP avait une longue tradition de recherches sur la synthèse à formants. Grâce aux liens étroits avec le KTH de Stockholm, au soutien financier du CNET, au travail imposant de Pierre Badin et de Gérardo Murillo, le laboratoire disposait d'un large corpus de trames paramétriques obtenues par analyse-synthèse de signaux naturels. Par édition interactive de trajectoires d'une vingtaine de paramètres formantiques décrivant l'enveloppe spectrale et de paramètres décrivant les diverses excitations, ces trames délivrées toutes les 5ms à un synthétiseur à formants généraient un signal synthétique quasiment indiscernable de l'original. En 1988, cet effort collectif nous a permis de donner vie à un premier synthétiseur à partir du texte utilisant une synthèse par concaténation de trames formantiques [bailly-etal:fase88] : approche assez paradoxale dans la mesure où la dissection du signal par synthèse à formants est souvent un préalable à une synthèse par règles. Les thèses d'O. Al Dakkak et M. Guerti [guerti:these93] y ont été d'ailleurs dédiées.

Construire une description « en compréhension » d'un ensemble de réalisations décrites « en extension » n'est pas une tâche aisée : dans le domaine de la synthèse à formants, mises à part l'approche par schémas prônée relativement récemment par J. Coleman & J. Local [coleman:tm92] [local:94] et la concaténation de blocks élémentaires (ABU pour « Acoustic Building Units ») utilisée dans MultiVoc [olaszy-etal:tm92], la démarche traditionnelle est une description de l'évolution des paramètres par des trajectoires connectant des points cibles. L'identification, la variabilité contextuelle voire l'existence de cibles formantiques constituent en eux-mêmes un espace toujours vivace de recherche en parole : cibles virtuelles, trajectoires soumises à réduction, espace formantique qui respire au rythme de l'information linguistique disponible a priori…. simple conséquence de la programmation de points d'équilibre musculaires sous-jacents sont autant de propositions où les recherches en synthèse de parole ont souvent tenu un rôle central ; d'une part pour leur propre besoin de modèles de description de la variabilité observée, d'autre part parce que les systèmes de synthèse fournissent un accès efficace à des stimuli contrôlés pour des expériences de perception et enfin parce qu'ils imposent un modèle d'analyse délivrant des trajectoires cohérentes.

Les premières parties des thèses de M. Guerti [guerti:these93] et d'A. Neagu [neagu:these98] ont été dédiés à un recueil de données formantiques sur les cibles des voyelles et consonnes du français et sur la caractérisation de leur variabilité contextuelle.

Formants et résonances des voyelles

Les cibles formantiques des voyelles varient suivant de nombreuses dimensions incluant le milieu de propagation, le locuteur, la situation de communication, le débit, la structure accentuelle et le contexte phonétique - dimensions d'ailleurs non forcément orthogonales. L'utilisation de corpus contrôlés autorise une étude relativement homogène des modes de variations suivant l'une de ces dimensions tout en conservant les autres conditions d'élocution

relativement stables. De nombreuses études ont ainsi été menées sur l'influence du contexte phonétique immédiat sur les cibles atteintes afin d'établir des modèles de coarticulation incluant une mesure de la capacité de chaque phonème à anticiper ou à préserver les caractéristiques phonétiques de ces voisins.

Dans la perspective d'obtenir des gabarits formantiques pour chaque voyelle de la banque de données formantiques, Guerti montre non seulement que toutes les voyelles ne sont pas également sensibles au contexte mais que tous les formants ne présentent pas la même variabilité. Ces « macro-sensibilités » acoustiques différenciées des voyelles s'expliqueront de trois manières :

(a) perceptive : les voyelles ouvertes ont des ellipses de dispersion qui semblent plus grandes que les voyelles fermées sur une échelle linéaire des fréquences mais relativement équivalentes sur des échelles logarithmiques

(b) articulatoire : les voyelles avec appui des articulateurs (langue, lèvres) semblent permettre un positionnement plus précis, plus stable des articulateurs. De même, ces positions induisent une saturation des recrutements musculaires.

(c) acoustique : les lois de l'acoustique et la géométrie spécifique du conduit vocal imposent une morphologie spécifique au contenu spectral des signaux de parole.

En effet, les nomogrammes de Fant ainsi que les simulations acoustiques utilisant des modèles articulatoires plus réalistes montrent que l'espace acoustique produit par un conduit vocal est fortement structuré et que les formants ne l'organisent pas suivant autant de dimensions indépendantes. La ré-interprétation des nomogrammes de Fant par Badin et al [badin-etal:jasa90] montrent que les formants peuvent être affiliés de manière majoritaire à une des cavités ou constrictions réalisées dans le conduit vocal par les articulateurs. Les formants s'organisent ainsi en familles de résonances organisées suivant les lois de l'acoustique en fréquence de Helmholtz, multiples de quart d'onde ou de demi-onde.

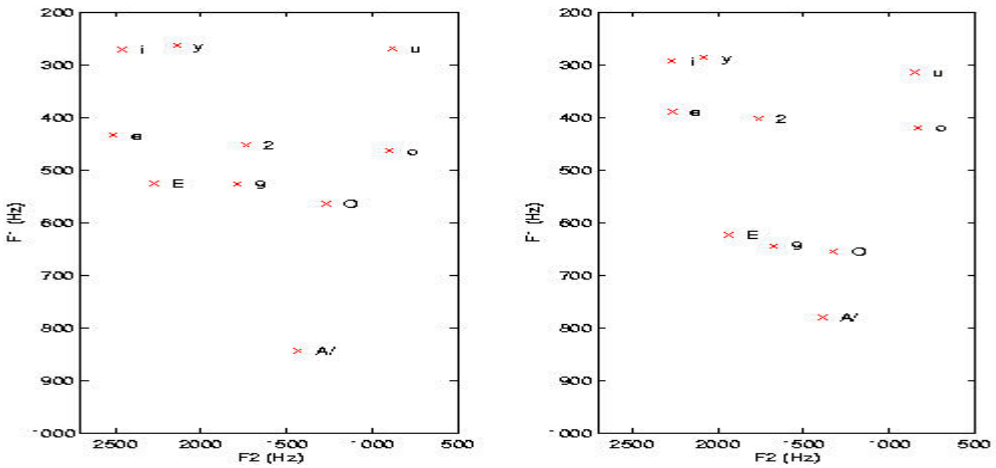

Figure 4: Triangles vocaliques (plan F1-F2) de deux locutrices étudiées par A. Neagu. On voit les diverses stratégies de répartition des voyelles intermédiaires dans le triangle.

Etude des affiliations par simulation articulatoire

Les nomogrammes de Fant utilisent cependant un modèle simplifié de la fonction d'aire [fant:1960] commandé par deux constrictions. Badin et al montrent que l'ensemble des maxima de la fonction de transfert de telles fonctions d'aires peut être approximé par les résonances de la cavité arrière + constriction linguale, de la cavité avant + constriction labiale et de la constriction linguale elle-même (étant fixée à 4cm chez Fant, elle a une demi-onde assez basse).

Pour des fonctions d'aire réelles, il est souvent difficile de déterminer l'emplacement optimal de la constriction linguale ainsi que l'affiliation majoritaire des formants: on utilise le plus souvent les macro-sensibilités qui consistent à perturber systématiquement l'aire de chaque section et d'interpréter ainsi comment les formants du conduit y réagissent. Fréquences de Helmholtz, multiples de quart d'onde ou de demi-onde ont alors un comportement spécifique (sensibilité à la constriction, phénomènes oscillatoires…) qui mettent en évidence des affiliations à des cavités identifiables.

De manière complémentaire à la méthode des macro-sensibilités, nous avons développé les nomogrammes dits à cavités découplées [bailly:levels95] qui consistent à couper de manière systématique le conduit en deux sous-conduits et en évaluant la capacité de l'ensemble des résonances de ces deux sous-conduit à prédire les résonances du conduit original. La coupure optimale délimite alors deux cavités et délivre comme sous-produit l'affiliation majoritaire de chaque résonance du conduit. Nous nommerons par la suite R2i-1 les résonances affiliées à la cavité arrière et R2i celles affiliées à la cavité avant.

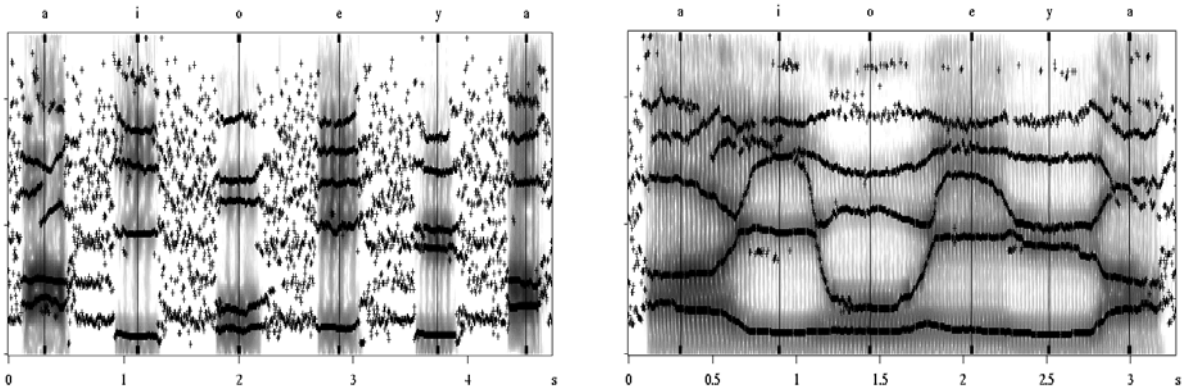

Figure 5 : résonances vocaliques. A gauche, la suite de voyelles [aioeya] prononcées isolément; à droite: en continu.

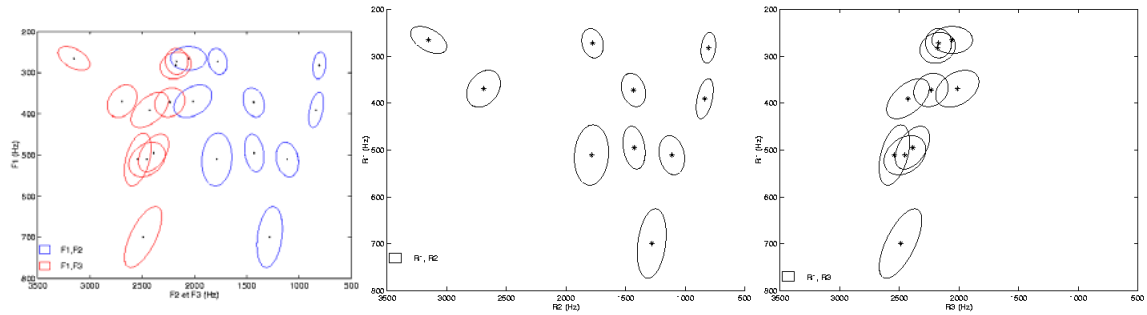

Figure 6: A gauche, les espaces (F1-F2) et (F1-F3) superposés. Au centre, l'espace (R1-R2) et à droite, l'espace (R1-R3).

Structuration de l'espace acoustique

Il restait à vérifier ces prédictions sur des signaux naturels de manière plus systématique que les premières propositions avancées par G. Kuhn [kuhn:jasa75] [kuhn:jasa79] puis par Hermansky [hermansky-broad:icassp92] sur les liens entre la résonance R2 et F'2 ne l'avaient fait. Nous avons donc étudié sur de nombreux locuteurs les transitions voyelle-voyelle, puis voyelle-occlusive-voyelle enfin les successions de trois voyelles.

C'est grâce au travail de M. Guerti, de L. Roussarie et d'A. Neagu que nous avons proposé une structuration de l'espace acoustique des voyelles en résonances [bailly:levels95] permettant ainsi de projeter l'espace F1-F2-F3 sur un espace à deux dimensions R1-R2 où les

cibles des voyelles sont bien séparées (voir Figure 6). C'est ainsi que nous avons montré que la résonance R3 pouvait être « éliminée » par filtrage adaptatif. Cette élimination semble plus rentable pour une classification catégorielle qu'une intégration large bande telle que suggérée par Schwartz et Escudier [schwartz-escudier:sc89] notamment pour expliquer le contraste [e] vs [ø].

Les trajectoires de résonances dans des continuums vocaliques peuvent être décrites comme le déplacement d'une particule entre points attracteurs « acoustiques » de manière analogue à ce que suggérait Ohman pour l'évolution de la coupe sagittale de son geste vocalique. C. Leclerc confirmera cette stratégie sur un corpus incluant des suites co-articulées de trois voyelles V1V2V1 : son travail montre que la réduction s'opère sur des trajectoires quasi rectilignes ouvrant ainsi la voie à une récupération « cinématique » des cibles intentionnelles : les trajectoires dans le plan R1-R2 « pointent » vers V2 même lorsque la cible est non atteinte. Cette réduction de dynamique semble donc réfuter l'hypothèse d'une « respiration » de l'espace vocalique dont les dimensions varieraient selon les exigences de clarté dans le modèle Hypo-Hyper de Lindblom [lindblom-lindgren:perilus85] [lindblom:icphs87] : les cibles semblent ici invariantes et exercer un pouvoir attracteur plus ou moins fort sur la trajectoire, modèle que nous reprendrons ci-dessous pour contrôler les trajectoires articulatoires par inversion articulatori-acoustique.

Un résultat intéressant du travail d'A. Neagu [neagu-bailly:jep96] sur la structuration de l'espace acoustique des voyelles concerne les voyelles intermédiaires : l'aperture des voyelles intermédiaires ne semble pas diviser l'aperture maximale de manière régulière (voir Figure 4). Diverses stratégies sont employées par les divers locuteurs pour remplir l'espace maximal : si certains locuteurs respectent la répartition logique 1/3-2/3 pour le français qui dispose de deux niveaux d'aperture intermédiaires, d'autres réalisent une partition 1/4-1/2 ou symétriquement 1/2-3/4. Il est à noter cependant que dans tous les cas toutes les voyelles intermédiaires d'un même niveau d'aperture sont alignées et présentent un premier formant centré autour de la même valeur : ce qui confirme bien une gestion de cette structuration par la mâchoire porteuse.

Identification des occlusives et résonances

Les voyelles constituent le point d'ancrage privilégié des modèles de coarticulation : d'après Ohman [ohman:jasa67], les gestes consonantiques rapides se superposent au geste vocalique lent, chaque cible vocalique étant, comme ci haut, décrite comme un barycentre des trois voyelles extrêmes. La lecture du spectrogramme va donc des voyelles aux consonnes. La théorie de l'invariance relative [sussman-etal:jasa91] [sussman-etal:jasa93] propose d'utiliser l'abaque F2 du centre de la voyelle versus F2 à l'attaque resp. à la fin de celle-ci. Chaque consonne décrit alors sur cet abaque un lieu spécifique permettant d'identifier la consonne. Ces lieux sont souvent caractérisés par des droites dites équations de locus. Si cette caractérisation cinématique permet d'obtenir de meilleurs taux de reconnaissance qu'une caractérisation plus statique des cibles consonantiques [nossair-zahorian:jasa91], elle est cependant contestée par des expériences de dévoilement progressif. Ces expériences montrent que les auditeurs peuvent identifier la consonne en se contentant de quelques millisecondes « contextuelles » du signal de la voyelle, en tout cas sans disposer de la cible vocalique. Les approches dynamiques disposant de l'évolution du spectre sur quelques dizaines de millisecondes à l'attaque de la voyelle (nossair) atteignent les meilleures performances disponibles à ce jour sur ces « vocabulaires » consonne-voyelle réputés difficiles.

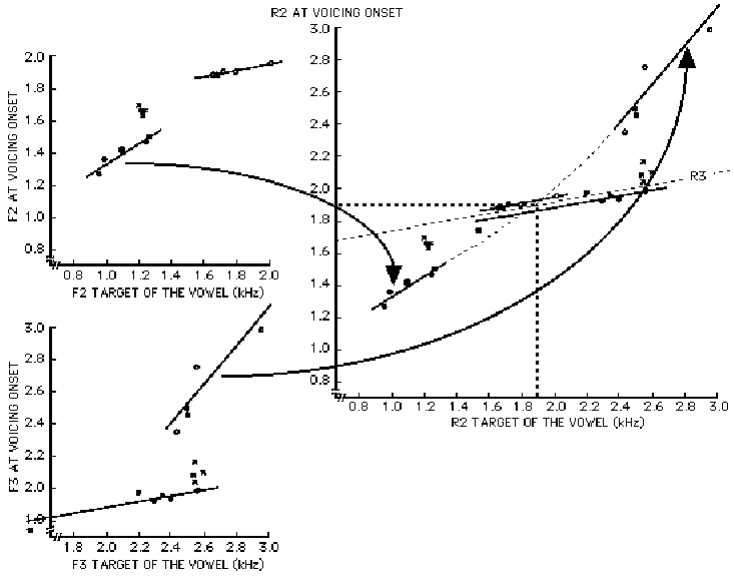

Figure 7: Superposition des équations de locus du [g] en F2 et F3 mettant en évidence un changement d'affiliation sous-jacent.

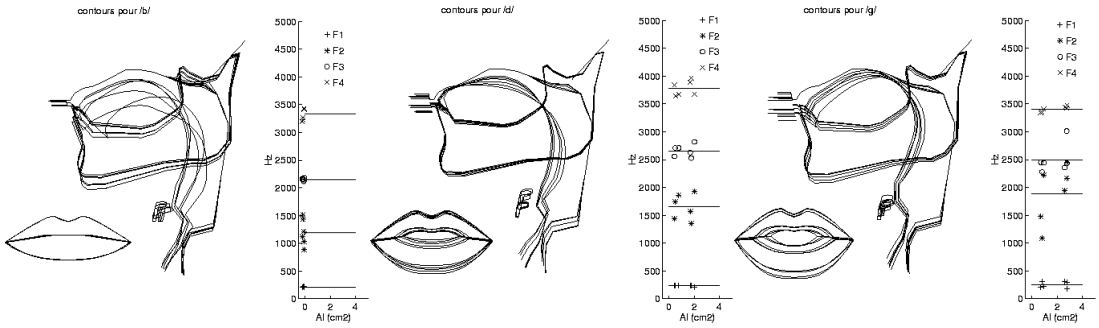

Figure 8: Simulation d'occlusions coarticulées. De gauche à droite : [b], [d] et [g]. On remarque que [g] présente une cible en F2 autour de 1500 Hz pour les faibles aires aux lèvres et une cible en F3 vers 3000 Hz pour des ouverstures plus importantes.

De la linéarité des équations de locus

La superposition des mouvements supposée linéaire par Ohman produit bien des équations de locus linéaires pour /b/ et /d/ alors que /g/ semble posséder deux lieux décrits comme deux « comportements » différents des deux allophones palatal et vélaire. Les simulations que nous avons effectuées (voir Figure 8) confirme le comportement non linéaire de la structure formantique de /g/. La superposition des abaques en F2 et F3 publiées par Klatt montrent cependant qu'il n'en est rien : les points d'intersection entre les deux lieux correspondent à un changement d'affiliation des formants considérés. Les abaques en R2 montrent ainsi des comportements sans ruptures et quasi linéaires.

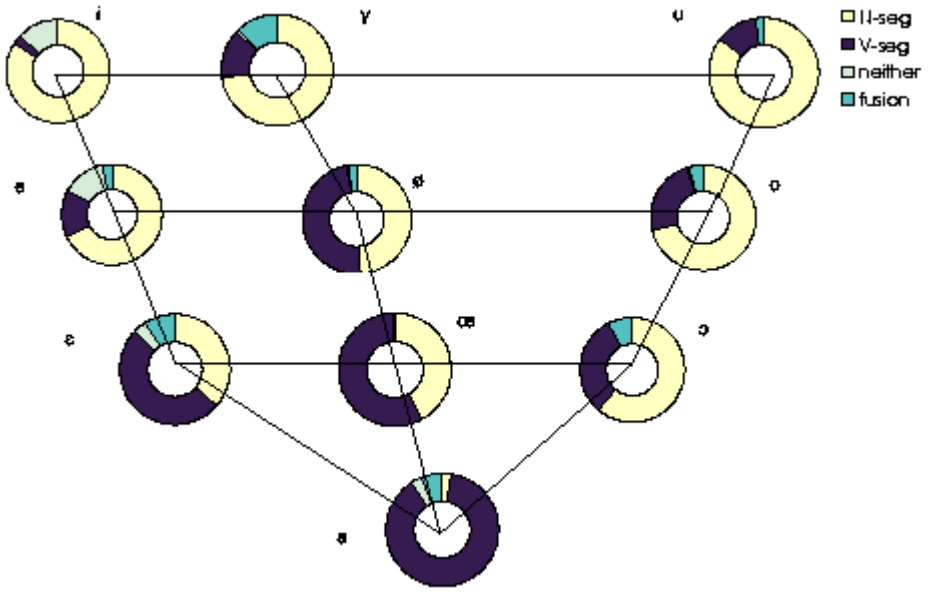

Figure 9 : Perception de stimuli conflictuels [neagu:these98].Influence du contexte vocalique sur le poids relatif des segments. On voit que le phonème indiqué par le segment vocalique (transitions des formants) l'emporte d'autant plus sur celui indiqué par le segment sourd (plosion + bruit de relâchement) pour les voyelles ouvertes que la voyelle est ouverte.

Cependant si ce ré-étiquetage permet de s'affranchir d'une rupture de comportement - ce qui a grandement facilité l'écriture de règles de coarticulation lors de la description des trajectoires formantiques entreprise par M. Guerti - , il reste que les lieux des diverses consonnes ne sont pas disjoints. Les équations de locus peuvent bien contribuer à démasquer certaines paires minimales mais elles ne le font pas seules. Smits et al [smits-etal:jasa96a] [smits-etal:jasa96b] montrent d'ailleurs que les meilleurs jeux de paramètres permettant de distinguer les plosives en contexte intègrent caractéristiques de l'explosion, des transitions formantiques et du VOT.

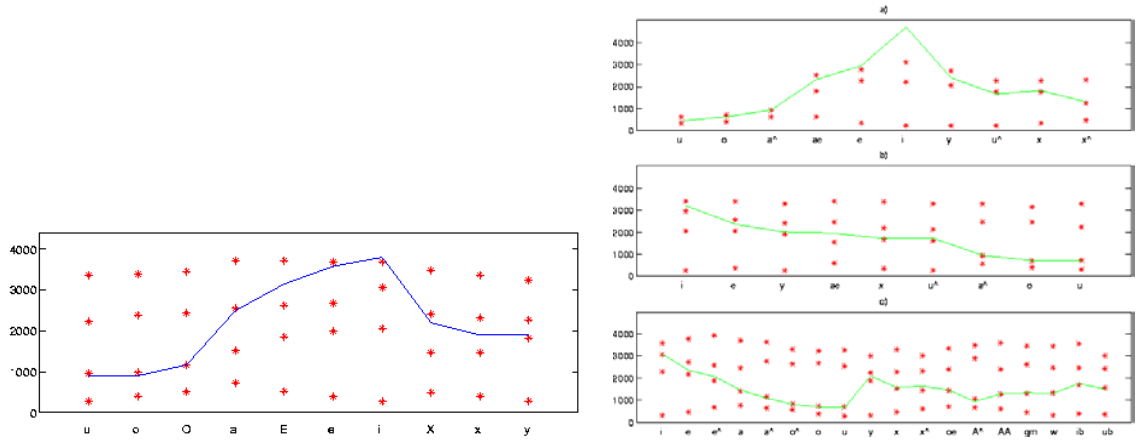

Figure 10 : A gauche, l'évoluation de la fréquence caractéristique de burst en fonction de la voyelle support. A rapprocher des données de F'2 rapportées dans la littérature à droite (a: Frant & Risberg, b) Carlson et al.c) Bladon & Fant).

Compétition et collaboration

De multiples expériences montrent d'ailleurs un mécanisme de compétition entre ceux deux composantes où le spectre de l'explosion est plus discriminant que les transitions dans un contexte de voyelle fermée alors que le contraire est vrai pour les voyelles ouvertes.

« Discriminant » ne veut pas dire « caractéristique » car les expériences de perception d'explosions isolées [bonneau-etal:jasa96] et de dévoilement progressif montrent que le spectre de l'explosion en contexte fermé et avant est peu discriminant en isolé et que beaucoup plus de signal vocalique est requis pour atteindre les performances d'identification des autres contextes [kewley-port-etal:jasa83].

Grâce à de multiples expériences de perception sur un inventaire vocalique beaucoup plus riche que la majorité des études réalisées jusqu'alors, A. Neagu montre que si le mécanisme de compétition est largement dominant (voir Figure 9), il coexiste avec un mécanisme de collaboration plus précoce où le contexte vocalique « guide » l'analyse spectrale de l'explosion en sélectionnant la partie la plus informative du spectre. Une fois encore, cette analyse semble être conditionnée par un suivi de résonances ou tout au moins par un calage de l'analyse sur une partie du spectre d'explosion, ce calage étant mieux compris en termes de résonances. Notons de plus que cette caractérisation est conditionnée par l'entropie et non l'énergie, l'absence d'énergie dans certaines zones pouvant très bien être porteuse d'information. On rejoint ici les problèmes des relations complexes entretenues entre traitement préalable du signal et traitement morphologique.

Traitements précoces vs tardifs

A. Neagu a systématiquement testé diverses architectures de reconnaissance à l'aide d'un modèle simple (analyse discriminante et noyaux gaussiens). Pour étayer le mécanisme principal de compétition mis en relief dans les expériences de perception, il montre que toutes les architectures fusionnant les informations sur les transitions formantiques et sur le contenu spectral de l'explosion réalisent un meilleur score de reconnaissance que les identifications séparées. Cependant cette fusion est d'autant plus efficace qu'elle survient précocement dans le traitement du signal. Ainsi parmi les algorithmes classiques, c'est la représentation intégrée de Nossair et Zoharian [nossair-zahorian:jasa91] qui réalise le meilleur taux de reconnaissance. Le traitement morphologique proposé atteint à peu près les mêmes scores mais avec 3 fois moins de paramètres.

Nous retiendrons de cette série d'expériences qu'un gros effort de travail sur les représentations phonétiques des sons en contexte peut encore être effectué de manière à guider les algorithmes probabilistes vers des solutions plus compactes et plus discriminantes. Nos travaux vont dans le sens d'une « analyse de scène » acoustique plus poussée où les représentations du signal sont adaptées au type d'allophone à identifier.

Contrôle sensori-moteur de l'articulation

Si le suivi en ligne des résonances du conduit vocal peut simplifier la tâche de représentation et d'identification des sons en contexte, il peut aussi contribuer à simplifier la représentation et le contrôle des mouvements en supprimant des non-linéarités dans la correspondance entre geste et son. Ce suivi peut être basé sur un système de perception aussi bien guidé par une connaissance préalable des propriétés du système de production que par l'observation par ce dernier de régularités sur les signaux. Il est difficile de séparer les contributions des systèmes purement ascendants détectant des évènements saillants ou des régularités dans les signaux des contributions des projections de nos attentes, de nos connaissances a priori sur la structure de ces signaux, acquises et construites. Cette construction coopérative des représentations morphologiques devrait s'illustrer dans la dynamique de l'apprentissage, comme le montre la plasticité du babillage canonique au langage maternel.

Il semble cependant raisonnable que le langage articulé exploite de multiples représentations du mouvement et de ces conséquences aérodynamiques et acoustiques. Les représentations corticales sont alimentées en continu de signaux issus de multiples capteurs

sur le mouvement - copie efférente, proprioception, capteurs de contact, de pression, d'écoulement, boucle auditive. Tout comme les différents dispositifs expérimentaux - cinéradiographie, IRM, articulographie, aérodynamique, électromyographie, enregistrements acoustiques… - qui permettent d'acquérir de nombreux signaux sur l'organe en mouvement ou le cerveau en action, ces capteurs délivrent des informations bruitées, incomplètes et à diverses échelles de temps et de précision sur l'activité langagière. D'où l'idée d'un modèle interne mettant en cohérence ces signaux et les conditionnant, que pratiquement tous les modèles de contrôle du mouvement postulent.

Cette abstraction de l'organe contrôlé peut alors servir de représentation au mouvement imaginé, à la préparation du mouvement réel et à sa programmation et sa régulation en ligne.

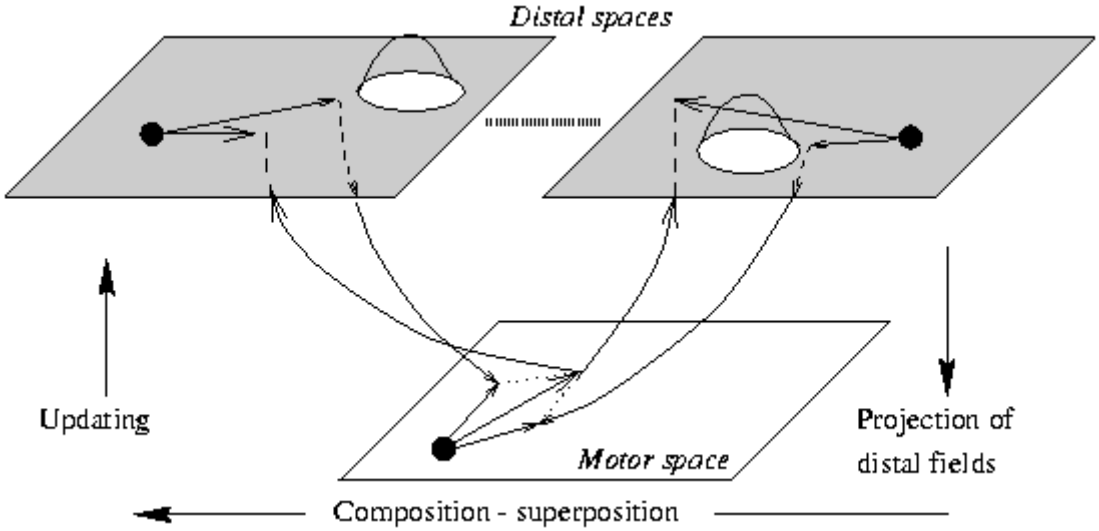

Figure 11: contrôle de l'articulation par cibles sensori-motrices. Les attracteurs sont activés dans des espaces concurentiels et de manière chevauchante (d'après [bailly:scom98][bailly:scom98]).

Inversion

Si la théorie motrice de la perception développée aux laboratoires Haskins s'appuie sur un modèle quantitatif de contrôle de la parole proposé dans le cadre de la phonologie articulatoire [browman-goldstein:85] [browman-goldstein:py86] [browman-goldstein:phono89], sa proposition théorique forte ne peut se targuer d'un appui quantitatif aussi fort en perception « inverse » : l'inversion articulatori-acoustique en est encore à ses balbutiements malgré un effort important d'évaluation [mcgowan:scom94] [saerens-etal:icphs91] [sorokin:scom92] [sorokin:scom94] [lin-fant:eurospeech89]. Il va de soi que l'inversion doit être fondée sur un modèle articulatoire fidèle, complet et performant intégrant une modélisation fine de tous les phénomènes aéroacoustiques et biomécaniques caractérisant le système de production. Cet effet énorme de collection de données, de modélisation physique et de simulation informatique ne peut être mené qu'au sein de large équipes pluri-disciplinaires.

Nous avons pour notre part identifié très tôt l'intérêt d'une régularisation de l'analyse acoustique par des contraintes articulatoires tant au niveau fréquentiel que temporel. Avec J.P. Liu [bailly-liu:ja90], nous avions étudié la possibilité de récupérer la structure formantique par projection sur le signal d'un modèle de Markov caché et avions conclu sur la perspective de récupérer ainsi des données sur le mouvement. La non disponibilité de larges bases de données articulatoires nous a empêché de continuer sur cette voie à l'époque. De même nous avions avec P.F. Marteau [marteau-etal:icassp88] [bailly-etal:icassp89] appliqué

le modèle de coarticulation de S. Ohman [ohman:jasa67], précurseur du modèle «Task dynamics», à un modèle de décomposition temporelle de manière à déconvoluer gestes vocaliques et consonantiques. Bien que ces travaux ne faisaient pas appel de manière explicite à un modèle articulatoire (non disponible à cette époque), ils constituent néanmoins un substrat de travaux dans lequel le travail de M. Jordan sur l'inversion de systèmes à degrés de liberté en excès [jordan:coins88] [jordan:attention90] [jordan-rumelhart:91] a pu trouver un écho immédiat [bailly-etal:asa90b] [bailly-etal:nsi90]. Grâce à l'expérience acquise, nous avons pu étendre la simple inversion de trajectoires acoustiques sous contraintes [bailly-etal:jphon91] à une inversion d'une tâche définie de manière plus composite par des gabarits multiparamétriques [mawass-etal:aa]. La monnaie d'échange entre auditeur et locuteur ne serait ainsi pas uniquement acoustique, géométrique ni articulatoire mais tout à la fois. Cette proposition rejoint la théorie de la perception directe proposée par C. Fowler [fowler:jphon83] [fowler:haskins90] [fowler:jasa91] [fowler:jasa96].

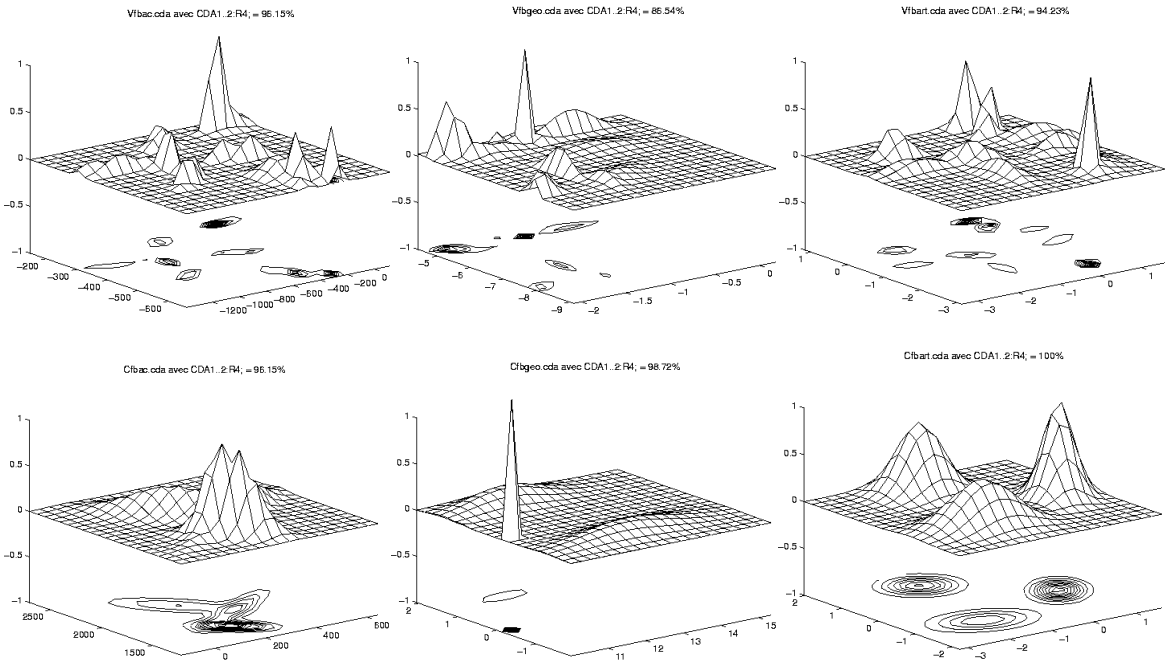

Figure 12 : espaces sensori-moteurs. De gauche à droite : l'espace acoustique (3 premiers formants), l'espace géométrique (constrictions) et articulatoire (7 paramètres). En haut: 10 voyelles; en bas, 3 occlusives. Les voyelles sont mieux séparées en acoustique, les consonnes en géométrie.

Représentations sensori-motrices

L'apprentissage d'un système phonologique particulier revient alors à sélectionner les gestes propres à reproduire au mieux les sons/gestes visibles d'un langage cible. Le modèle proposé par Markey [markey:these94] est à ce titre très séduisant : il envisage un contrôle en boucle où des unités de programmation - très semblables à des demi-syllabes - sont considérées atteintes et ainsi enchaînées dès lors qu'une partition acoustique propre à chaque unité est jugée exécutée. Le modèle de contrôle agit dans un premier temps de manière passive (voir le « passive motion paradigm » de Morasso & Sanginetti) mais collecte de manière collatérale l'ensemble des signaux accessibles au cours du mouvement. Ce qui lui permet alors de sélectionner parmi ceux-ci, dans un deuxième temps, les signaux de contrôle lui permettant de contrôler en ligne au mieux l'exécution correcte du mouvement. Cette sélection a posteriori des « meilleurs » paramètres de contrôle s'oppose au choix délibéré fait dans la plupart des autres modèles où l'encodage de la tâche phonologique utilise souvent une

partition neuromusculaire (ex : théorie du point d'équilibre), gestuelle [guenther:these92] [guenther:pr95], une partition acoustique [bailly-etal:jphon91] mais rarement une combinaison et encore moins une stratégie opportuniste comme le suggère Markey [markey:these94].

Si nous avons dans un premier temps privilégié une partition acoustique [bailly-etal:jphon91] mettant en valeur le rôle prépondérant de l'acoustique dans l'apprentissage du langage, nous avons récemment proposé un contrôle sensori-moteur de l'articulation [bailly:scom98] qui peut être qualifié d'opportuniste dans le sens où nous supposons que ce sont les représentations les plus discriminantes d'une classe de sons qui sont préférentiellement recrutées pour définir la partition : la représentation ponctuelle de l'articulation courante se déplace ainsi dans un champ de forces sensori-motrices modulées par des attracteurs activés de manière séquentielle mais chevauchante (voir Figure 11). Ce travail s'est largement appuyé sur le projet Speech Maps, dont j'ai coordonné pendant 3 ans les activités dans le domaine du contrôle. Cette activité se poursuit en collaboration avec D. Beautemps dont le projet de recherches vise à mieux définir l'incidence des variations des contraintes linguistiques et environnementales sur l'articulation.

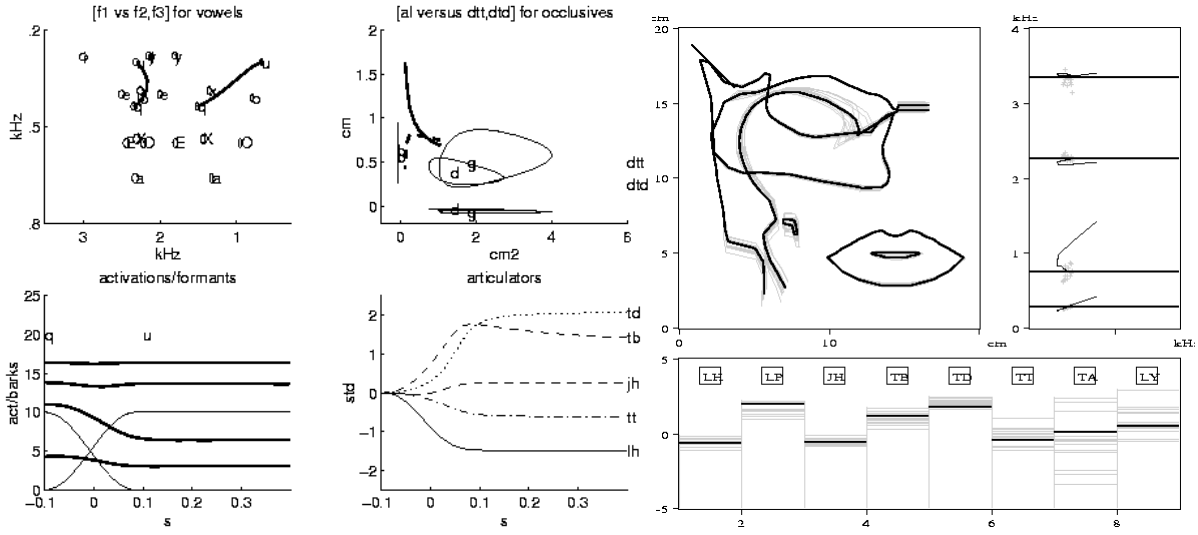

Figure 13 : activation de la cible du [u] depuis une configuration articulatoire neutre. A gauche, les trajectoires sensori-motrices décrites. A droite, comparaison avec les cibles vocaliques extraites de la base de données rayonsX [badin-etal:ica95].

Le contrôle sensori-moteur de l'articulation proposé procède donc d'un processus de sélection sur des représentations sensori-motrices du corps en mouvement, agissant sur et réagissant à l'environnement. Ce processus de sélection serait guidé par une exigence d'émergence et de compacité de cibles permettant l'étiquetage de cibles phonologiques et de construction du lexique mental (voir Figure 12). Le lien entre cette navigation virtuelle et la nécessaire prise en compte de la réalité biomécanique des organes contrôlés a fait l'objet d'une discussion dans un article publié dans le Bulletin de la Communication Parlée en commentaire d'un article de Perrier et al [perrier-etal:jshr96], où est esquissé un modèle de contrôle de l'articulation par point d'équilibre. J'y reprend l'idée de Mussa-Ivaldi et Bizzi [mussa-ivaldi-bizzi:ms97] qui proposent de coupler le mécanisme d'apprentissage cinématique du « passive motion paradigm » avec un mécanisme d'apprentissage dynamique par mise en forme d'une dynamique corticale. En effet les champs agissant dans les cartes sensori-motrices peuvent être mise en forme de manière à compenser - ou à exploiter - la dynamique propre au système biomécanique contrôlé de manière que les trajectoires réalisées dévient de manière minimale par rapport aux trajectoires virtuelles planifiées. Les trajectoires

réalisées pourront ainsi obéir à des contraintes - cinématiques *et* dynamiques - non seulement imposées par le système biomécanique mais les contraintes du système complet d'encodage de l'information incluant à la fois les exigences de robustesse au milieu d'exécution et de perception du geste et de catégorisation souhaitée par le système de signes sous-jacents.

« *Chez le (sujet) normal l'objet est « parlant » et significatif, l'arrangement des couleurs « veut dire » d'emblée quelque chose, tandis que chez le malade la signification doit être apportée d'ailleurs par un véritable acte d'interprétation.* »
[merleau-ponty:45] *p.153*

STRUCTURATION PROSODIQUE

La « prosodie articulatoire » est un domaine de recherches émergent qui s'intéresse justement aux stratégies articulatoires déployées pour structurer le message de manière à construire, mettre à jour « l'espace de croyance mutuel » entre interlocuteurs et répondre aux variations des conditions de communication. La compréhension des mécanismes de structuration prosodique du message est non seulement un enjeu majeur de la contribution des sciences de parole aux sciences de la cognition mais répond aussi à un défi technologique majeur : véhiculer/extraire du sens non littéral par/depuis un ensemble de signaux sensibles. Par la parole, nous véhiculons non seulement le contenu informationnel du simple énoncé mais aussi tant d'autres informations sur le locuteur, sur son état physique et émotionnel, sa position sur son propre discours [bolinger:89] ainsi que sur les conditions environnementales de l'articulation. Selon le modèle Hyper/hypo proposé par Lindblom et collègues [lindblom-lindgren:perilus85] [lindblom:icphs87], ces tâches sont de plus négociées avec l'interlocuteur selon la croyance que nous avons des connaissances a priori que nous lui supposons.

S'il est difficile de désolidariser le sens de son substrat comme le montre bien le débat entre Changeux et Connes [changeux-connes:92][changeux-connes:92][changeux-connes:92] sur les rapports entre neurobiologie, neurophysiologie et mathématiques, nous nous sommes essentiellement intéressés aux structures sonores dont l'intonation pouvait disposer pour véhiculer du sens de manière parallèle, parfois concomitante aux agents linguistiques dont l'interlocuteur dispose pour interpréter notre discours.

Avant de procéder plus avant et d'illustrer ce que j'entends par structures sonores, il me semble important de mettre en exergue les rapports que peuvent entretenir à mon sens structuration prosodique - donc structures sonores -, intonation et paradigme morphologique.

Prosodie et morphologie

Bien qu'il soit difficile démontrer l'apport d'intelligibilité au message dû de manière intrinsèque à l'intonation sans avoir recours à des artifices souvent peu écologiques, cet apport est incontestable, ne serait-ce que par sa fonction de démarrage « bootstraping » de l'apprentissage de la structure morpho-syntaxique de la langue [morgan-demuth:96]. Cette fonction de démarrage s'illustre aussi dans les expériences de transcription de phrases sémantiquement imprédictibles [benoit-etal:scom96] qui montrent que l'intonation permet de fournir un découpage préalable du message en unités de sens, des points d'ancrage que nos connaissances linguistiques peuvent alors travailler et affiner. Ces tests élémentaires d'intelligibilité, combinés à des mesures de temps de réaction ou à des tâches annexes de mémorisation, permettent de comparer l'efficacité de diverses stratégies d'encodage prosodique de l'information linguistique voire para-linguistique y compris celles mises en œuvre dans des synthétiseurs de parole. Les mesures de temps de réaction ou de dérive des scores en fonction d'un accroissement de la tâche permettent de quantifier la charge cognitive de la tâche principale et ainsi des ressources sollicitées par l'auditeur pour assurer un décodage suffisant du message.

Les tests d'opinion « Mean Opinion Scores » sont aussi largement utilisés pour faire évaluer ce surcroît de charge de manière consciente par l'auditeur. Ces tests couplés au

précédents permettent non seulement de savoir quelles sont les parties de discours mal « interprétées » - au sens mal jouées - mais de raffiner plus avant ce jugement afin d'en connaître les causes. Il est en effet important de discerner les causes multiples de cette « mésinterprétation » de la partition informationnelle : est-elle due à une mésentente sur la tâche ou plutôt sur la manière de la nommer - donc sur un espace de croyance mutuelle mal estimé par l'auditeur -, à un choix possible mais inapproprié de la structure intonative en regard de la tâche, à un choix erroné mais faisant partie des inventaires intonatifs de la langue ou enfin à une structure intonative inacceptable.

Cette notion de percept, de forme bien formée est à la base du Gestaltisme. L'un des pères fondateurs de la théorie, Max Wertheimer, a essayé d'établir expérimentalement l'existence d'une « bonne forme ». Si les exemples privilégiés des gestaltistes appartiennent aux illusions optico-géométriques et au domaine des figures réversibles, on peut néanmoins trouver des exemples dans le domaine acoustique : une mélodie n'est pas saisie comme une série de notes distinctes, mais comme un tout que l'on peut transposer d'un ton dans un autre ton ; cette mélodie sera cependant radicalement modifiée si on modifie une ou deux notes, un ou deux éléments rythmiques.

Les travaux de J. Piaget montreront cependant que certaines propriétés des formes sont invariables, ou à peu près, tout au long de l'existence humaine, tandis que d'autres se modifient au cours de la vie. Cette notion de bonne ou mauvaise forme pourrait ainsi être le résultat d'une complexe composition de propriétés physiques inaltérables du monde qui nous entoure - ex : gravité, persistance des objets -, de propriétés de nos organes de production et de perception - ex : résolution, couplages perceptuo-moteurs [viviani-stucchi:advpsy92] - et les significations associées aux formes et induites par son environnement.

Il est donc concevable que les formes perçues puissent être plus ou moins facilement interprétées en fonction de leur adéquation à cet ensemble de propriétés plus ou moins élaborées, plus ou moins immédiatement accessibles à l'analyse. Dans l'ensemble des arguments que l'on peut développer en regard de cette perception orientée formes plus en rapport avec le domaine acoustique, citons l'illusion de Pogendorf répliquée dans le domaine acoustique [kluender:perilus91] montrant notre capacité à évaluer la « rigidité » d'une modulation de fréquence, et la théorie de l'attente en perception musicale. Boltz & Jones montrent effectivement l'existence de « ryhthmical and melodic expectation » sur des schémas musicaux élémentaires tronqués, ceci sur des sujets musiciens et non-musiciens [jones-boltz:psyr89][boltz:rpp92]. De même, F. Grosjean [grosjean:ling83] [grosjean-hirt:lcp96] montre que nous pouvons estimer la longueur d'un d'énoncé tronqué.

Ces expériences montrent notre capacité à projeter sur le continuum acoustique des attentes (des contraintes ?) sur le décodage à venir d'unités sonores et ceci à divers niveaux de résolution et de manière adaptative. A notre sens, ces attentes participent à un processus qui permet de déconvoluer les structures informationnelles du discours projetées sur un même continuum sonore suivant des dimensions non orthogonales (contrairement à ce que laisse supposer M. Rossi [rossi:speechcom93] sur la congruence entre structure prosodique et informationnelle). Ce processus d'analyse-synthèse permet ainsi de « suivre » en ligne plusieurs structures informationnelles en les actualisant suivant les déviations observées entre formes attendues et observées.

Il est donc capital de comprendre quelles sont les propriétés des formes intonatives bien formées depuis les propriétés de structuration du signal jusqu'aux formes plus larges d'encodage de l'information linguistique. Nous insisterons par la suite sur la robustesse de l'encodage spatio-temporel de l'information et essayerons de montrer que l'information doit être et est encodée de manière redondante et répartie.

Structuration rythmique

Notre activité est conditionnée par de nombreux rythmes biophysiques comme le rythme cosmique ou le rythme cellulaire ou de rythmes biologiques comme la respiration, la digestion, les phases du sommeil ou le cycle menstruel. Ces rythmes correspondent à la récurrence d'un repère physique ou d'un événement à intervalles temporels plus ou moins réguliers. Ces multiples rythmes permettent à notre organisme de réguler son activité et de maintenir un équilibre entre ses diverses composantes. Ces divers rythmes sont donc synchronisés et toute perturbation d'un des rythmes entraîne des perturbations dans la structure rythmique générale.

Outre le fait que la parole partage l'organe de la respiration et de l'alimentation, la structuration rythmique d'un message procède de multiples propriétés de l'organisation spatio-temporelle des sons. La récurrence de nombreux phénomènes sonores en parole telle la présence régulière de transitoires entre sons, l'alternance consonne-voyelle, la présence d'accents, de pauses, de formes intonatives récurrentes… peut être le support de la perception de multiples rythmicités.

Une rythmicité « bien formée » de ces évènements doit donc permettre une synchronisation des divers traitements de la parole, linguistiques aussi bien que pragmatiques incluant la gestion harmonieuse ou agressive des tours de parole ! Gérer de manière optimale cette rythmicité (ou plutôt ces rythmicités) est un enjeu majeur des systèmes de synthèse. Cette gestion optimale peut cependant être obtenue de deux manières diamétralement opposées :

- soit en se fixant comme objectif la réplication la plus minutieuse possible de la séquence de sons observée lors de la production d'un message répondant à la tâche fixée par un locuteur humain - donc en respectant de manière implicite ces multiples contraintes rythmiques . L'objectif est alors ici l'erreur de prédiction nulle des paramètres prosodiques en termes objectifs.
- soit en essayant d'identifier et de caractériser ces diverses composantes rythmiques, notamment les repères qui les ancrent dans le substrat sonore et la manière dont ces repères se synchronisent lors de perturbations.

Bien sûr, cette dichotomie est toute théorique, quoique dépeignant bien le paysage des recherches en synthèse de parole : l'identification de composantes rythmiques essentielles peut effectivement permettre de pondérer les erreurs objectives et de « bons » systèmes de prédiction peuvent constituer un outil remarquable d'expérimentation pour la modélisation.

Structuration rythmique par prédiction des durées segmentales

Un système permettant de prédire, à partir de facteurs phonotactiques, linguistiques ou paralinguistiques, les durées phonémiques - en supposant que les repères rythmiques soient effectivement calculables à partir d'un positionnement correct des instants de plus grande instabilité spectrale - et la structure des autres évènements prosodiques contribuant à la structuration rythmique, et ceci avec une erreur nulle, résoudrait de manière implicite la gestion rythmique du message.

La minimisation de l'erreur de prédiction des durées segmentales par un modèle paramétrique se heurte cependant à de nombreuses difficultés. De nombreux facteurs influencent ces durées, outre les nombreux facteurs disons « prosodiques », de nombreux facteurs « intrinsèques » tels que l'aperture, le voisement de l'allophone, « co-intrinsèques » tels que la nature des phonèmes adjacents voire phonotactiques tels que le nombre de phonèmes de la syllabe, la longueur du mot… viennent modifier la durée nécessaire à la réalisation souhaitée du phonème. Une étude expérimentale exhaustive de ces facteurs est quasiment impossible de part le nombre et l'orthogonalité de ceux-ci. Ainsi même si on peut

envisager de construire des corpus ad hoc - voir par exemple une étude des durées des logatomes desquels sont extraits la plupart des unités minimales utilisées par les systèmes de synthèse par concaténation, ces données constitueront de toutes façons une population peu dense (voir d'ailleurs ce problème évoqué par van Santen [vansanten:atr96]. Le recours à des modèles quantitatifs imposant une structure à la fonctionnelle f telle que d=f(Pi) permet de restreindre l'espace des relations possibles.

Ces fonctionnelles sont des fonctions simples : affines, polynomiales, … où des facteurs sont ajoutés, multipliés en fonction d'un arbre de décision portant sur des traits extraits de la structure phonologique de l'énoncé : nature des constituants, contexte phonétique, nombre de constituants de la syllabe, position dans la syllabe, position de la syllabe dans le mot… Ces facteurs permettent de factoriser le comportement de nombreux segments et ainsi d'augmenter la représentativité par des connaissances a priori. Ainsi de nombreux modèles additifs [oshaughnessy:jasa84], multiplicatifs [oshaughnessy:jphon81] et plus récemment de sommes de produits [vansanten:tm92] ont été proposés. van Santen souligne l'importance du traitement statistique a priori [vansanten:sc92] [vansanten:csl94] permettant de dégager les traits importants, de les ordonner et d'étudier la nature de leurs interactions afin d'éviter à nouveau une explosion combinatoire de la composition des traits. Quel que soit le soin apporté aux choix des prédicteurs, le but fixé à ces systèmes reste la minimisation de l'erreur de prédiction des durées segmentales.

Or cette minimisation conduit parfois à des résultats décevants d'un point de vue perceptif : le résultat de la minimisation objective de l'erreur de prédiction des durées segmentales n'est pas forcément préféré [tournemire:these99] à une modélisation basée sur l'expertise. Nous pensons que ceci peut être expliqué par le fait que l'organisation temporelle d'un énoncé n'est pas le produit d'une simple juxtaposition de segments phonétiques mais le résultat de la superposition de *structures rythmiques* de plus haut niveau possédant chacune une structure interne cohérente, de *flux rythmiques* plus larges qui se synchronisent pour accomplir un certain nombre de tâches dont celle de permettre la production/perception - parfois juste suffisante - des unités segmentales. L'ancrage acoustique de ces structures rythmiques ne nécessite pas une reproduction fidèle de tous les évènements saillants accessibles à l'analyse.

Structuration rythmique par ancrage syllabique

De nombreuses études montrent que la syllabe est une structure centrale de la programmation motrice. L'importance de l'oscillation mandibulaire dans l'émergence du langage est illustrée par les données sur le babillage canonique et la structure des premiers mots [davis-macneilage:ls94] qui montrent une structure CVCV majoritaire et un système d'association préférentielle entre certaines consonnes et voyelles. Konopovsky propose que la structure rythmique des langues se construit progressivement à partir de cette isochronie initiale et intègre progressivement des caractéristiques propres à la langue et ceci de manière d'autant plus difficilement que la structure rythmique de la langue s'éloigne d'une isochronie syllabique.

D'autres études ont montré que des auditeurs pouvaient capturer de manière cohérente et reproductible le rythme syllabique et le synchroniser avec d'autres signaux périodiques ou d'autres activités rythmiques telles que frapper du doigt, de la main et du pied. C'est dans l'hypothèse d'une synchronisation de flux par détection d'évènements sur les signaux que Marcus [marcus:these76] a proposé le Perceptual Centre ou le moment d'occurrence perceptif de la syllabe. Les données de production et de perception accumulées depuis par Marcus, Pompino-Marshall [pompino-marschall:jphon89], Scott [scott:these93] situent cet instant au voisinage de l'établissement acoustique de la voyelle.

La structuration rythmique par l'organisation temporelle des moments d'occurrence perceptif permet alors de concevoir une prédiction des durées segmentales en deux étapes : un positionnement des moments d'occurrence perceptifs fixant ainsi des rendez-vous temporels entre rythme syllabique et organisation segmentale des syllabes (au niveau des établissements vocaliques). Un premier modèle de ce type a été élaboré par Campbell et Isard dans un cadre purement syllabique [campbell-isard:jphon91]. Ils montrent qu'un modèle de répartition de la durée syllabique exploitant un modèle très simple d'élasticité des segments permettait d'expliquer la variance des durées segmentales observées sur un large corpus. C'est dans un cadre analogue que Plinio Barbosa a proposé et étudié la structure rythmique des Groupes Inter -Perceptual Centre (GIPC) en fixant non plus comme points d'ancrage de la structure rythmique les établissements syllabiques mais les établissements vocaliques.

Il va de soi que cette gestion de flux rythmiques peut être étendue à d'autres niveaux de programmation du langage sur lesquels peuvent peser d'autres contraintes de production ou de perception (par exemple couplage entre unités propositionnelles, respiration et phonation).

Contours rythmiques

On peut opposer à l'argumentation précédente le fait que les systèmes de prédiction des durées segmentales intègrent de manière implicite de nombreuses caractéristiques syllabiques telles que composition ou position de la syllabe courante dans l'énoncé : la médiation syllabique peut alors émerger de la procédure d'apprentissage du modèle de prédiction. Or cette médiation n'a pas comme raison essentielle d'ajouter une contrainte a priori supplémentaire. Elle constitue surtout la clé d'une analyse rythmique réelle : elle donne accès au flux rythmique via un marquage acoustique : elle permet d'avoir accès et d'étudier la structure rythmique comme une courbe temporelle comparable à la mélodie. Cette médiation met en forme les signaux de segmentation de manière à étudier leur régularité temporelle, leur rythmicité.

En Français, le groupe accentuel est décrit comme l'ensemble des syllabes inaccentuées précédant une proéminence rythmique et mélodique appelée Accent Final (AF). La synchronisation entre structure mélodique et rythmique s'effectue sur cette position accentuelle qui reçoit un contour mélodique en fonction de la fonction du groupe accentuel dans le discours. A cette définition du groupe accentuel minimal est souvent ajouté la présence éventuelle intermédiaire d'un accent dit secondaire, dont la fonction n'est pas phrastique. Mertens y ajoute un appendice subséquent et facultatif de syllabes atones.

Plinio Barbosa a étudié l'évolution temporelle des intervalles inter-Perceptual Centre - ou durée des GIPC - :en Français : les maxima de cette courbe correspond bien à ce qui est décrit généralement comme l'accent final. La grande majorité des courbes entre ces maxima consécutifs évolue de manière monotone et asymétrique : l'allure générale est ascendante. Le groupe accentuel est ainsi caractérisé par un ralentissement progressif du rythme syllabique qui culmine à l'AF. Le groupe accentuel suivant débute alors par une « remise à zéro » (reseting) du rythme syllabique. Cette étude confirme les données de Padeloup sur la durée syllabique.

A notre sens, ce contour rythmique du groupe accentuel en Français n'est pas forcément accessible à une prédiction directe des durées segmentales où tous les prédicteurs de la durée seraient mis au même niveau. Nous avons montré que les auditeurs sont beaucoup plus sensibles à une perturbation de cette forme particulière du contour rythmique du groupe accentuel en Français qu'à une perturbation équivalente apportée de manière plus uniforme sur les durées segmentales : la forme fait donc sens au-delà de ses caractéristiques les plus saillantes. Le contour rythmique est ainsi insécable et participe dans sa globalité dans l'accès à la fonction de l'unité accentuelle. Il permet ainsi de délimiter l'unité : notamment en caractérisant le regroupement à gauche ou à droite des syllabes attenantes au fait saillant (voir

Figure 14). Il permet de plus à l'interlocuteur d'anticiper la venue d'un fait saillant et d'estimer la distance au fait saillant [grosjean:ling83]. Cette conception de la structuration rythmique par régulation d'un flux rejoint la théorie de l'attente rythmique de Jones [jones-boltz:psyr89] [boltz:rpp92]: les auditeurs ont une attente manifeste sur la structure rythmique d'un morceau tronqué. Le fait que cette attente semble relativement indépendante de l'expérience musicale des auditeurs montre que cette faculté perceptive est relativement universelle et « bas-niveau » dans la structuration temporelle du flux sonore.

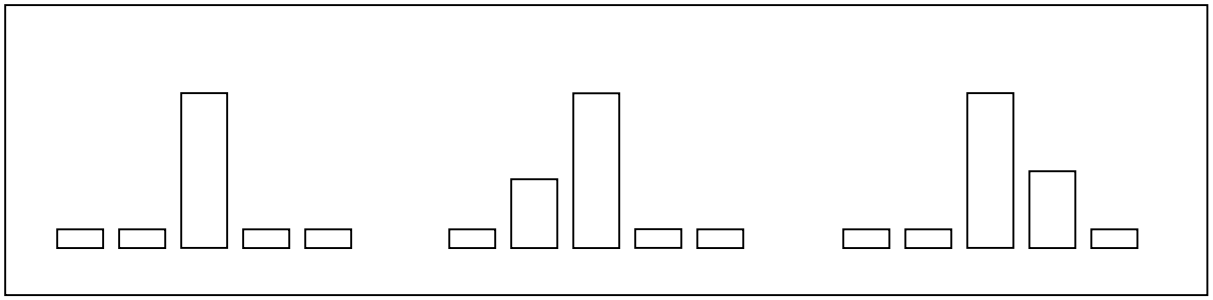

Figure 14: Shémas rythmiques. A gauche, rien ne permet d'indiquer la structuration du groupe rythmique. Au centre, un groupe « trail-timed », terminé par l'accent, et à droite « head-timed », commençant par l'accent.

Elle relève pour nous de ce niveau morphologique, où le code exploite une structuration « naturelle » du monde physique : la continuité rythmique exprime la cohésion naturelle d'un phénomène oscillatoire non entretenu, une discontinuité étant interprétée comme la conséquence d'une commande intentionnelle.

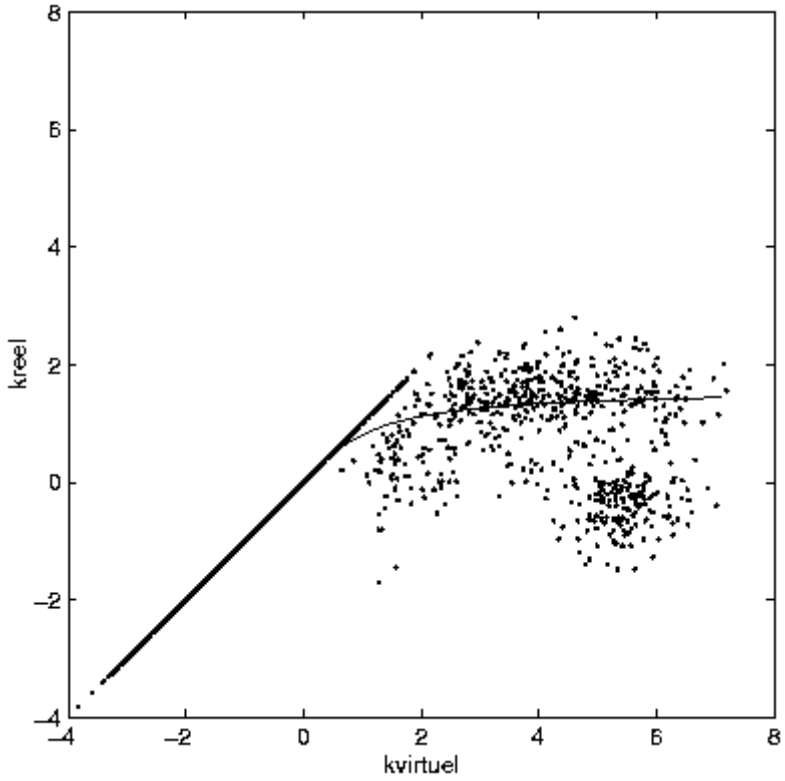

Figure 15 : émergence de la pause silencieuse. Les points représentent des GIPC prélevés sur le corpus "Formules" de Bleike Holm [holm:these]. En abscisse; l'allongement virtuel (sans pause); en ordonnée, l'allongement de la partie sonore seule. La courbe en trait continu représente la loi utilisée d'émergence utilisée en synthèse: la différence entre la courbe et la droite kreel=kvirtuel donne détermine la durée de la pause. On voit une zone de chevauchement autour de kvirtuel=2 entre stratégies pause subjective (fort allongement) versus

pause silencieuse. Au delà d'un allongement de 4, l'allongement virtuel peut revenir à 0, ce qui est souvent le cas en fin de phrase.

Emergence de la pause

La pause joue un rôle très central dans la plupart des modèles intonatifs et rythmiques. En phonologie prosodique, elle délimite des unités phonologiques de rang supérieur appelées « Intonation Units » ou « Intonation Phrases ». Ce segment phonétique jouit donc d'un statut phonologique particulier. Il nous a toujours semblé très surprenant que la pause en tant que phénomène émergent, lié au débit et dépendant dans une moindre mesure d'une stratégie de marquage propre au style de parole et au locuteur, conditionne de manière aussi massive la structure sonore profonde. P. Barbosa a examiné de manière plus attentive le mécanisme d'émergence de la pause en demandant à un locuteur de prononcer une même phrase à plusieurs débits. Nous avons à cette occasion proposé une méthode originale de contrôle du débit : au lieu de donner au locuteur une horloge de référence de manière explicite via un métronome par exemple, nous le lui avons donnée via une question synthétique dont le débit est contrôlé de manière explicite : le locuteur devait alors répondre à cette question suivant le même tempo que cette dernière. L'échange d'horloge est alors implicite et les mesures de débit montrent a posteriori que cet échange s'est effectué au mieux car nous avons ainsi obtenu 5 débits clairement différenciés.

Cette étude montre que l'émergence de la pause est progressive passant par une étape de type « catastrophique », de changement de phase (voir Figure 15) où le locuteur peut choisir entre un allongement syllabique excessif et la génération d'une courte pause (jusqu'au 50ms[3]). Nous avons proposé ainsi un modèle d'émergence de la pause intégré dans un modèle de répartition des durées segmentales à l'intérieur du cadre de programmation rythmique par GIPC présenté ci-dessus (voir article joint au dossier).

Les termes de la négociation entre rythme et contenu segmental

L'oscillation mandibulaire « naturelle » régie par la fréquence propre du système hyoïdo-mandibulaire n'est pas seulement perturbée par la structuration rythmique de haut niveau, signalant ainsi par des déviations d'une oscillation purement isochrone la présence d'une information intéressante. Le temps disponible doit aussi permettre d'articuler et de percevoir les sons et il n'est pas étonnant que le substrat sonore impose des contraintes à la programmation rythmique.

Les diverses mesures de durées que nous avons effectuées sur de multiples corpus confirment les tendances montrées par Fant et Kruckenberg [fant-kruckenberg:icslp96] dans le cadre syllabique dans de nombreuses langues : la durée d'une unité de discours quelconque semble être globalement proportionnelle au nombre des unités qu'elle contient même si une faible tendance à la compression est attestée pour de longues unités. L'isochronisme syllabique de certaines langues [dauer:jphon83] ne serait donc juste qu'un artefact phonotactique et résulterait de la structure syllabique la plus fréquente.

Les premiers corpus que nous avions étudiés étaient construits avec des syllabes de type CV afin de pouvoir utiliser au mieux les résultats de psychophysique et de perception du rythme : peu d'études ont prolongé les travaux sur le P-Center avec des attaques et coda consonantiques plus complexes ni des syllabes en séquences. Il reste que ce travail reste à mener soit en production en imposant une rythmicité (métronome) extérieure ou en demandant à réitérer des énoncés avec des contenus phonétiques différents (voir par exemple

[3] Il est cependant difficile d'étiqueter précisément l'instant de disparition d'une structure sonore car l'énergie s'affaiblit de manière continue et les pauses internes entre segments voisés s'accompagnent souvent du maintien d'un faible murmure glottal.

les travaux en réitération d'énoncés entrepris par A. Rilliard), soit en perception en continuant ces expériences d'ajustement.

Perspectives

Les contraintes pesant sur la rythmicité mandibulaire semblent nombreuses et leur négociation adaptative. Ceci expliquerait pourquoi les énoncés supposés être prononcés avec un rythme régulier et identique avec une large variété de contenu phonétique (notamment les logatomes posés dans une phrase porteuse permettant d'extraire des polysons) présentent un allongement syllabique très marqué en fonction du nombre de constituants de la syllabe. En effet, la prononciation de logatomes est une situation où le contenu phonétique est peut prédictible et la structuration linguistique très pauvre. C'est donc l'organisation temporelle précise de l'articulation qui l'emporte en l'absence de contraintes d'encodage de structuration linguistique. A l'inverse on peut imaginer des situations où le discours possède une structuration linguistique complexe. A cet égard l'étude réalisée par H.R. Pfitzinger [pfitzinger:icslp98] montre que le tempo ne peut être estimé qu'en prenant en compte à la fois le nombre de phonèmes ET de syllabes par secondes. Dans une expérience originale, il a demandé à des auditeurs de classer des morceaux de signaux de 625ms extraits d'un corpus de dialogues simulés *PhonDatII* lus par 16 locuteurs différents suivant le tempo perçu : les auditeurs manipulaient ainsi des petits symboles correspondant aux énoncés sur un écran d'ordinateur, le placement en abscisse correspondant au tempo perçu par rapport à trois items de référence (lent, normal et rapide). Ils pouvaient les écouter autant qu'ils voulaient en cliquant dessus. L'expérience dure en moyenne une heure et les auditeurs n'éprouvent pas de difficulté à classer les stimuli de manière relative et convergent vers des résultats très comparables (r=0.96). Le classement moyen présente une corrélation moins élevée avec le nombre de phonèmes (r=0.73) qu'avec le nombre de syllabes (r=0.81) prononcées par seconde alors que la corrélation atteint 0.96 en utilisant les deux prédicteurs. Il serait ainsi intéressant de reproduire cette expérience avec des énoncés mieux contrôlés sur le plan de la richesse phonotactique et du contenu informationnel.

Structuration intonative

Tout article ou livre sur l'intonation commence par un panégyrique des fonctions remplies par l'intonation dans la structuration du discours. Segmentation, hiérarchisation et focalisation d'unités linguistiques, véhicule essentiel des attitudes et des émotions, porteuse de l'identité, du sexe, du poids (!), de l'appartenance socio-culturelle, de l'état physique du locuteur, l'intonation véhicule tant de canaux d'information sur le locuteur, le contenu linguistique et la situation de la communication qu'on est souvent étonné par la pauvreté des systèmes de notation prosodique de l'intonation. La tentation symbolique prend en effet ici tout son sens. Comment en effet décrire un paysage complexe à l'aide d'un vocabulaire d'une dizaine de mots ? Bien sûr, cette description utilisera le mot « arbre », distinguera éventuellement les sapins des bouleaux, les mots « ciel », « nuage », distinguera la falaise du ravin et le torrent du ruisseau. Une description plus informative ajoutera des termes configurationels indiquant l'emplacement relatif et mettant en relation de proximité ou de dépendance ces diverses unités de signification. Rien cependant ne garantit que le lecteur ou l'auditeur saura dessiner ainsi le paysage à l'identique ni même garder son caractère. Les systèmes de marquage prosodique - ce langage sur le langage - n'exploitent en effet typiquement qu'une dizaine de symboles factorisant l'ensemble des faits saillants observés dans le continuum prosodique : tons, accents sont alors organisés dans une structure phonologique les mettant en relation syntaxique voire sémantique [marsi-etal:tm97]. Charge alors à l'analyse du signal et à la reconnaissance des formes de mettre en relation le continuum prosodique et ces objets phonologiques, et aux systèmes de traitement symbolique

de modéliser les relations entretenues entre ces objets et les autres agents de communication du langage.

Analyse ascendante

Ainsi une analyse purement ascendante cherchant à extraire l'information saillante offre l'avantage de ne pas « filtrer » les observations par un modèle fonctionnel donné a priori. De nombreuses méthodes de stylisation semi-automatique du continuum prosodique ont été proposées, principalement dédiées à la mélodie : segmentation en séquences de tons syllabiques statiques ou dynamiques [allessandro-mertens:csl95], de contours syllabiques [emerard-etal:tm92], modèles de proéminence accentuelle [taylor-black:etrw94] [dusterhoff-black:etrw97], modèles de cibles connectées [pierrehumbert:jphon90] [pierrehumbert:jasa81] ou de commandes [fujisaki-sudo:areri71] [fujisaki-kawai:icassp88]… L'idée directrice est toujours de coller au plus juste à l'observation au regard de distances objectives intégrant des contraintes de perception (égalité perceptive du modèle IPO) ou de production (fréquence de coupure dans le modèle de Fujisaki ou négligence des erreurs négatives dues à la micromélodie dans Momel).

Ce premier niveau garantit souvent la préservation de la structuration intonative du message sachant que cette vérification s'effectue par des systèmes d'analyse-synthèse où les durées naturelles sont conservées. En effet, en regard de ces stylisations mélodiques pléthoriques, peu de modèles de stylisation rythmique ont été proposés, sachant que ceux-ci se concentrent essentiellement sur la syllabe finale des mots.

L'émergence d'un niveau plus profond de description phonologique rend nécessaire un deuxième niveau de quantification de cette première stylisation. Ce deuxième niveau correspond à la notion d'équivalence perceptive du modèle IPO. Il revient dans tous les cas à associer à un fait observé ou à un ensemble de faits une fonction équivalente dans la langue. On retrouve ici le parallèle avec la structure segmentale et la distinction entre allophones et phonèmes : chaque fait prosodique est supposé être une réalisation, une instanciation d'une unité minimale distinctive.

On dispose ainsi de systèmes de transcription prosodique permettant la segmentation et l'étiquetage du continuum sonore en unités minimales de sens. Contrairement à INTSYN [hirst:cutler-ladd83] [hirst:icphs91] [hirst:lund93] ou au modèle IPO qui procèdent en respectant les deux étapes de stylisation décrites ci-dessus, ToBI (pour « TOnes and Break Indices) propose un système de transcription plus direct de l'intonation qui fait un large appel à l'expertise [silverman-etal:icslp92] [beckman-ayers:tobi94] [beckman-hirschberg:tobi94]. ToBI présente ainsi un grave biais méthodologique dû au fait que l'expert a accès tout à la fois à l'énoncé textuel, au message sonore et aux paramètres prosodiques. Certains contours mélodiques peuvent être ainsi « phonologiquement » identiques alors que leurs réalisations phonétiques diffèrent largement. De plus, ce système de transcription mêle des critères hétérogènes de segmentation et d'étiquetage : si les « Tones » font bien référence à des mouvements mélodiques, les « Break Indices » sont ici des valeurs quantifiées (variant de 0 à 4) indiquant un degré de cohésion entre deux mots consécutifs sans aucun rapport objectif immédiat avec la valeur de l'allongement final ou le degré de rupture mélodique :

« The break-index tier marks the prosodic grouping of the words in an utterance by labelling the end of each word for the subjective strength of its association with the next word, on a scale from 0 (for the strongest perceived conjoining) to 4 (for the most disjoint). These categories of association strength, or `break indices' are based on work by Mari Ostendorf, Patti Price, Stefanie Shattuck-Hufnagel, and their associates (see, e.g., Price et al., 1991)... The break indices are meant to be a label of the SUBJECTIVE strength of the boundary. » [beckman-ayers:tobi97]

ToBI ajoute ainsi de nombreux niveaux d'interprétation de la substance phonétique faisant un large appel à l'expertise et effectue une étroite quantification des faits émergents. Cet étiquetage ne garantit pas un quelconque niveau de transparence et ne décrit aucune méthode de resynthèse possible permettant de contrôler perceptivement l'expertise effectuée.

Analyse descendante

Comme souligné plus haut, les objets phonologiques émergents assurent diverses fonctions discursives : bien sûr, segmentation, hiérarchisation et mise en relief d'unités mais aussi « valence » informative de ces unités dans le système de croyance mutuelle où on trouve encodé ce que Bolinger appelle le positionnement du locuteur vis à vis de son propre discours [bolinger:89]: émotions, attitudes, modalités ainsi que de nombreux systèmes de signes sociolinguistiques et géo-culturels. Il est alors raisonnable d'envisager un cheminement inverse partant de la fonction à l'encodage, du sens au substrat physique. L'observation des données n'est alors plus liée à leur saillance immédiate mais à leur pouvoir explicatif, coextensif voire discriminant de la fonction qu'elles réalisent. Cette observation ne peut donc avoir lieu que dans un système plus global d'oppositions fonctionnelles : pour faire émerger le mécanisme d'encodage, il faut retrouver dans les données immédiates les indices permettant aux interlocuteurs d'échanger, de « pointer » une ou plusieurs fonctions dans l'ensemble des possibles. On comprend dès lors pourquoi une telle démarche est beaucoup plus marginale dans la littérature : on ne peut plus raisonner sur quelques échantillons de parole alors qu'ils assurent une multiplicité de fonctions discursives. Il faut donc observer d'importantes quantités de données, avoir recours à des corpus construits pour étudier quelques fonctions « in vitro » avant de les pister « in vivo »… et se doter d'un modèle apte à recevoir, à assumer les résultats de ces études de quelques échantillons de matière prosodique.

Avant de continuer, il est remarquable de noter qu'on retrouve dans les critiques adressées à cette approche expérimentale de l'intonation où la culture « in vitro » des fonctions intonatives est pratiquée, les mêmes reproches faits à la physique expérimentale du XIXème siècle. Alors qu'il semble tout à fait accepté que la matière ne livre sa profonde nature qu'enfermée dans des accélérateurs de particules, que l'infiniment petit n'est observable que s'il est « éclairé » par une lumière dont la longueur d'onde est de dimension comparable et donc d'énergies de plus en plus grandes., qu'une des meilleures façons de produire cette énergie est de construire des chocs entre particules… les corpus « in vitro » - où des consignes implicites ou explicites sont données au locuteur - sont souvent dénigrés et les résultats qu'ils délivrent jugés inexploitables pour analyser les corpus « in vivo » ou spontanés. A l'instar de la table de Champollion, les corpus « in vitro » sont pourtant la clé de l'accès aux procédures d'encodage des fonctions discursives qui nous permettront de construire les modèles - réellement explicatifs et non intuitifs - qui pourront alors être confrontés à la chimie du langage spontané.

Les chambres à bulle, à fils, à dérive ou à plasma, les calorimètres délivrent une multitude de données sur les collisions mais ces masses gigantesques de données seraient inintelligibles si on ne disposait que de modèles d'interprétation comme la théorie des forces régissant à divers niveaux de granularité la structure de notre univers. Bien que la simple mise en correspondance de données et de fonctions par des outils de la reconnaissance des formes (arbres de régressions multilinéaires, chaînes de Markov, réseaux de neurones,…) ait un succès croissant dans de multiples domaines : transcription orthographique-phonétique, génération des durées, cette analyse descendante simpliste a de même des limitations : la représentativité statistique des corpus doit garantir l'observation par les modèles de correspondance de toutes les combinaisons de facteurs influençant la structure prosodique - notamment phonotactiques, lexicaux, syntaxiques… l'explosion combinatoire suppose ainsi de disposer de larges bases de données étiquetées. Sans modèle d'interprétation donné a

priori, les outils mentionnés ci-dessus - qui opèrent par minimisation de l'erreur globale de prédiction des données - risquent fort d'opérer un lissage aveugle ou d'interpréter faussement des cooccurrences fortuites entre données et fonction comme la caractérisation forte de cette fonction. L'un des moyens d'éviter un tel biais est soit de construire de toutes pièces le corpus de manière à assurer une distribution optimale des facteurs soit de sélectionner de manière automatique ce corpus parmi une large base de données par des algorithmes « gourmand » ou génétiques [emerard :rhodes]. Un moyen encore plus sûr est d'observer les données via un filtre morphologique qui décrit le principe général d'encodage des fonctions linguistiques et paralinguistiques de l'intonation sur le continuum prosodique.

Une morphologie intonative

Bien qu'il soit difficile de séparer les contributions de l'organisation temporelle, de la mélodie et des variations d'intensité et de timbre à l'organisation rythmique et intonative de l'énoncé, la mélodie reste l'objet privilégié des études intonatives.

L'accent

Dans la plupart des modèles, les structurations rythmique et intonative « s'accordent » via la structure accentuelle. La structure accentuelle propose des rendez-vous entre celles-ci et permet ainsi d'assurer la synergie des faits prosodiques nécessaire à la perception correcte des dimensions perceptives. Peu de modèles proposent une réelle étude multi-paramétrique des paramètres prosodiques qui pourrait permettre de s'affranchir du fait accentuel lui-même pour ne retenir que sa fonction sous-jacente - morphologique, lexicale ou phrastique. Le fait accentuel, le fait de distinguer les syllabes accentuées des autres est en effet un filtre morphologique prégnant favorisant une analyse ascendante basée sur la proéminence : l'énoncé est alors fractionné en unités accentuelles et ne peut conserver sa cohésion, son unité fonctionnelle que grâce à une structure phonologique riche permettant de gérer les dépendances sémantiques à plus long terme des faits intonatifs - sans forcément y parvenir [marsi-etal:tm97]. La plupart des systèmes de description incorporent cependant des faits intonatifs qui échappent à la structuration accentuelle : une ligne de déclinaison basse, une ligne de déclinaison haute ou un mécanisme plus complexe, le « downstep » sont souvent gérés de manière parallèle [ladd:jasa88] et assurent des fonctions dialogiques universelles telles que la modalité ou la gestion du propos.

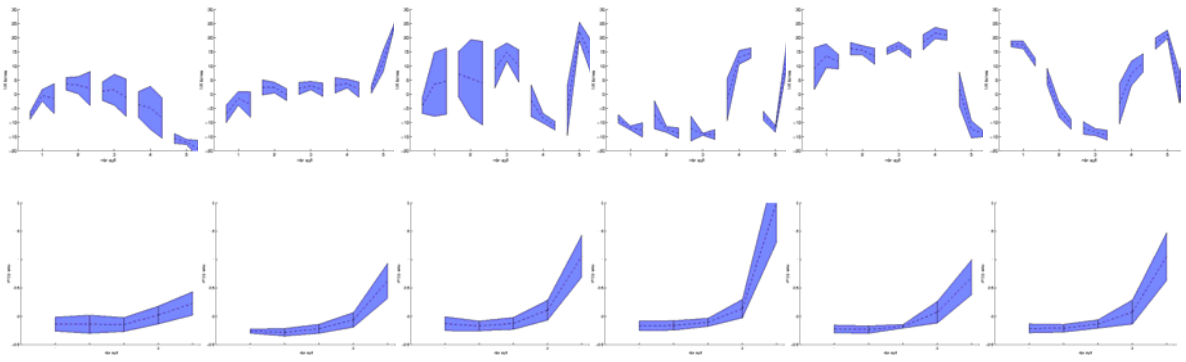

Figure 16: 6 attitudes prosodiques du Français: déclaration, question, exclamation, incrédulité, ironie de soupçon, évidence. En haut: contours mélodiques; en bas: contours rythmiques.

Une morphologie fonctionnelle

De manière opposée et complémentaire, V. Aubergé a proposé en 1991 [auberge:these91] [auberge:tm92] [auberge:lund93] une conception plus fonctionnelle, moins immédiatement

descriptive de l'intonation. Cette analyse bien que s'appuyant sur des travaux antérieurs [thorsen:fl83] [fonagy:82] [fonagy-etal:fl84] cède moins à l'intuition qu'à un paradigme théorique fort : un encodage de fonctions de structuration du discours par formes intonatives globales où pour ainsi dire « *la forme fait sens* ». Au delà d'un modèle complet de structuration de l'intonation que le lecteur pourra puiser dans ses différentes publications et dans le dossier d'habilitation qu'elle ne manquera pas d'écrire, c'est cette relation forte entre sens et substance, cet ancrage immédiat de la structuration qui m'a intéressé au prime abord. Si les conséquences en terme de « bootstrapping » de l'apprentissage de la structuration du langage et du discours par la prosodie sont apparues plus tard, ce modèle théorique prolonge par un ancrage cognitif plus marqué le simple ancrage supposé physiologique du modèle de Fujisaki que j'avais adopté pour ma thèse [bailly:these].

Le modèle de Fujisaki suppose en effet que le double contrôle indépendant exercé sur la fréquence fondamentale par la pression sub-glottique et la tension des cordes vocales est capturé, « phagocyté » par des fonctions phonologiques distinctes (resp. accentuation de groupe de sens et de souffle), effets alors superposés par simple addition des composantes. A ce contrôle « proximal » des degrés de liberté des effecteurs laryngés - similaire à l'exploitation de l'oscillation mandibulaire dans le langage articulé - peut alors se greffer un niveau plus « distal » de représentations guidées non plus par la seule physiologie mais par une nécessité de véhiculer un signifié plus riche et plus profond.

Le travail de Yann Morlec [morlec:these97] sur les attitudes prosodiques a démontré que cet encodage de fonctions par formes globales était opérationnel tant du point de vue de la description des contours prosodiques (voir Figure 16) que de sa validité comme monnaie d'échanges : en effet, l'encodage par contours globaux présente la propriété de répartir l'information, l'entropie du message sur une partie d'énoncé non limitée à un ensemble de syllabes proéminentes. C'est l'ensemble du contour qui participe à l'encodage et se différencie des autres contours encodant les interprétations concurrentes de la partie d'énoncé concernée. Les expériences de « gating » menées par T. Crépillat [auberge-etal:eurospeech97] sur ces attitudes et confirmées par d'autres chercheurs [vanheuven:etrw97] ont montré que les résultats d'identification de contours tronqués pouvaient être interprétés par un mécanisme de compétition, de comparaison dynamique du contour se déroulant dans le temps et un ensemble de contours référentiels permettant l'accès à des interprétations concurrentes et donc par l'existence de véritables « points d'unicité » des contours prosodiques référentiels.

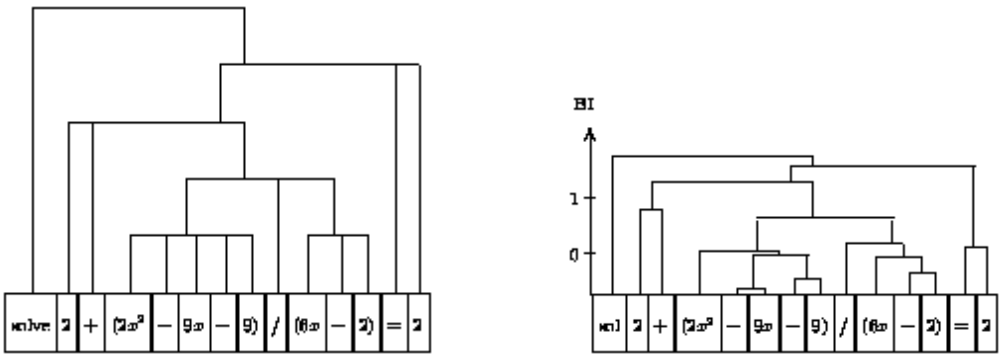

Figure 17 : structure de performance d'une formule mathématique énoncée. A gauche; l'arbre syntaxique; à droite, la structure de performance (d'après [holm-etal:icphs99]).

Nature et fonction des constituants

La deuxième proposition fondamentale de V. Aubergé et coextensive à la précédente est que les fonctions assumées par ces contours - que nous appelons aussi mouvements - prosodiques coextensifs à des unités de discours (syllabes, mots, groupes syntaxiques,

propositions, phrases…) doivent être véhiculées en coopération avec les autres agents linguistiques (lexique, morpho-syntaxe, sémantique…) de structuration du discours. On parlera de « rendez-vous » structurels. Ces rendez-vous sont à rapprocher de la notion de marqueurs prosodiques utilisés par une importante quantité de chercheurs français [emerard:these77] [dicristo:these85] mais jouissent d'une plus grande autonomie par rapport à la structure prosodique de surface, notamment par rapport à la notion d'accent. Ils n'appartiennent en fait à aucun niveau de description (prosodique, lexical, morpho-syntaxique, pragmatique…) particulier mais permettent de synchroniser l'activité de ces divers agents en des points particuliers ou sur des étendues identifiées du discours. Ils assurent donc des fonctions plus générales telles que la démarcation, la mise en relief, la mise en relation d'unités au sein du discours mais aussi le degré d'adhésion, la charge affective du locuteur à cet élément de discours.

Ainsi si les rendez-vous et contours du modèle initial dépendaient de la nature et de la fonction des unités considérées au sein de niveaux linguistiques très hiérarchisés et si le modèle actuel garde encore cette adéquation, cette mise en correspondance rigide entre niveaux linguistiques et fonctions, je pense que la prosodie assume les fonctions générales énoncées ci-dessus et ceci de manière relativement indépendante du niveau linguistique considéré : elle possède la faculté générale de mettre en relief une unité quelque soit sa nature et sa fonction au sein de la langue… mais la met évidemment le plus souvent au service de la structuration linguistique! C'est ce que nous testerons au cours de la thèse de B. Holm [holm:these] consacrée à l'étude de l'énonciation de formules mathématiques.

B. Holm a déjà montré que les structures de performance [holm-etal:icphs99] des énoncés peuvent refléter de manière très complète le riche degré d'enchâssement de la structure syntaxique des formules mathématiques (voir Figure 17). La question qui est alors posée est de savoir :

si cette riche structuration résulte de l'identification de plusieurs niveaux de structuration - à l'image d'une identification de propositions, syntagmes, de groupes, de sous-groupes et de mots - et véhiculant chaque niveau avec des mouvements prosodiques spécifiques,

ou si cette riche structuration résulte plus simplement d'une capacité plus générale à véhiculer des relations de présupposition, de hiérarchie d'unités de plus en plus larges. Cette structuration s'applique alors de manière récursive en utilisant – réutilisant - à chaque fois les mêmes mouvements.

Négociation structurelle

La combinaison la plus simple de mouvements élémentaires permettant de véhiculer la structure globale de l'énoncé est la superposition. Cette stratégie a été adoptée par de nombreux auteurs que se soit en prosodie ou de manière générale dans beaucoup de domaines du contrôle moteur. Cette superposition se résume le plus souvent à une simple opération de superposition/addition sans interaction des mouvements superposés.

Au cours de la thèse de Y. Morlec et dans les premières analyses de B. Holm, il est cependant manifeste que la complexité de certaines fonctions, la connaissance présupposée de certaines par l'interlocuteur, l'accroissement du nombre de rendez-vous et des diverses fonctions demandées à la prosodie influencent de manière manifeste l'encodage de certaines fonctions.

Y. Morlec avait proposé de contrôler de manière explicite l'amplitude de modulation exercée sur chaque paramètre prosodique par un mouvement élémentaire en fonction de l'importance relative des diverses fonctions exercées par chaque mouvement au sein de l'énoncé global. La richesse du canal de communication (signal acoustique, débit, présence conjoint du visage ou de bruit environnemental …) et de l'espace de croyance mutuel entre interlocuteurs (historique du dialogue, connaissances linguistiques et socioculturelles…)

présenterait ainsi une capacité limitée d'encodage dans laquelle le locuteur choisirait de clarifier, de mettre en valeur des fonctions plus que d'autres : on peut donc envisager de mettre en valeur la mise en valeur d'une partie de discours! De même la tendance à la compression de l'amplitude de modulations permettant de hiérarchiser de petites unités lors de leur enchâssement dans de plus larges formules a été mis en exergue par B. Holm.

Ce simple modèle ne permet cependant pas de rendre compte de manière satisfaisante de la totalité de la variabilité observée. Ceci peut être dû à plusieurs raisons. Chaque fonction ne correspond pas de manière univoque à un contour/mouvement prototypique mais à un ensemble de variantes, de configurations possibles jouant à la fois sur la configuration et sur le caractère multiparamétrique de l'encodage. Le modèle de négociation par contrôle de l'amplitude de modulation est de plus sûrement trop simpliste et ne prend pas en compte des stratégies locales et globales plus complexes résultant de problèmes tels que des conflits entre contours superposés.

CONCLUSIONS

Energie versus entropie

Le deuxième principe de la thermodynamique énonce que l'entropie d'un système isolé ne peut que croître. En conséquence, lorsqu'un système a atteint son état d'équilibre, son entropie est maximale. Ce terme est repris par la théorie de l'information pour quantifier la fiabilité d'un système à coder et transmettre des messages[4]. Dans ce bref survol des pistes de recherches que nous avons suivies, nous nous sommes attachés à montrer que l'accès à l'information par une chaîne de traitement procédant par schématisation, par « épouillages » successifs de données brutes par des filtres de plus en plus « haut niveau » doit incorporer une boucle d'analyse par la synthèse. Cette boucle permet à l'interlocuteur de projeter sur les divers niveaux de représentation des signaux (sensori-moteurs, phonologiques, linguistique, paralinguistiques et émotionnels) divers niveaux d'attente sur information qui lui permettront d'assurer une plus grande robustesse de la communication car ce mécanisme permet de restituer facilement des données manquantes (dues à un manque de clarté ou d'attention, ou à un environnement de communication perturbé) mais surtout de porter efficacement l'attention de l'interlocuteur sur des parties de discours par des déviations exercées sur des formes attendues.

Ainsi je pense que les systèmes sous-symboliques viennent généralement trop tard chercher l'invariance et modéliser la variabilité dans la chaîne de traitement alors que les premiers étages de traitement ont déjà centré l'analyse sur les parties souvent « énergétiques » des signaux de communication. Comme nous l'avons vu dans le cas de la perception des occlusives en contexte, du rythme ou de l'intonation, des « détails », des déviations subtiles sur les signaux peuvent être porteuses de sens dès lors que l'on recherche des fonctions de discrimination au sein d'un système.

Comme dans le cas de la thermodynamique nous dirons qu'un système de communication cherche à maximiser son entropie et que c'est tout autant la valeur informationnelle d'un indice que son énergie qui détermine son utilisation - son émergence - au sein du système.

Il va donc de soi que la communication s'opère ainsi au sein d'un système de conventions - de manière à ce que ces attentes soient semblables au sein d'une communauté linguistique donnée -. C'est la connaissance de ce système de conventions, de ses fondements cognitifs qui nous permettront d'étudier sur quel fond se meut l'interaction langagière. A cet effet les études expérimentales que nous menons où l'activité langagière est contrôlée et le système de conventions disséqué peut être considéré comme l'étape incontournable d'obtention des mécanismes de base d'encodage de l'information. Ce contrôle permet d'avoir les appuis statistiques suffisants pour valider les hypothèses cognitives sous-jacentes. Grâce à ces outils d'analyse et de modélisation, c'est alors - et seulement alors - que nous pourrons disséquer des

[4] Selon Shannon, l'information contenue dans un message est une quantité mathématiquement mesurable, liée à la probabilité que ce message soit choisi parmi un ensemble de messages possibles. Plus le message est probable, plus la quantité d'information qu'il transporte est faible. Par conséquent, un message attendu avec certitude possède une quantité d'information nulle. Dans la plupart des applications pratiques, lorsque l'on décide d'envoyer un message, on le choisit parmi un ensemble de messages possibles. Tous ces messages sont susceptibles d'être transmis, mais avec une probabilité qui leur est propre. On désigne alors par *entropie,* terme emprunté à la thermodynamique, la moyenne des quantités d'information des différents messages possibles.

interactions verbales plus libres, plus riches, plus complexes et que nous pourrons les situer dans des cadres connus et préalablement décrits.

Emergence des représentations

On l'a vu dans les travaux dans ce mémoire, le paradigme morphologique tend à laisser émerger les représentations intermédiaires sous l'action conjuguées de deux types de contraintes : des contraintes de bas-niveau, puisés dans les systèmes de production et de perception, et des contraintes de plus haut niveau, concernant la nature des informations que l'on cherche à véhiculer et la fonction qu'elle assure dans le système de communication. L'émergence de ces représentations intermédiaires est donc tributaire tout à la fois des représentations paramétriques des signaux (et ceci est relativement crucial dans le domaine de la caractérisation de l'articulation et de la prosodie articulatoire) et des modèles d'organisation phonologique que nous présupposerons. Le choix délibéré fait ici est de faire des hypothèses très minimalistes sur les principes d'organisation et de supposer que l'organisme procède par « greffage » d'un système de communication sur des aptitudes, des espaces catastrophiques préexistants, en faisant donc un appel minimal à l'arbitraire du signe et à l'intelligence calculatoire.

BIBLIOGRAPHIE

[abry-badin:etrw96] C. Abry & P. Badin
Speech mapping as a framework for an integrated approach to the sensori-motor foundations of language
ETRW on Speech Production Modelling: from Control Strategies to Acoustics, Autrans - France, 1996.

[allen:jphon75] G. D. Allen
Speech rhythm: its relation to performance universals and articulatory timing
Journal of Phonetics, 3:75--86, 1975.

[allessandro-mertens:csl95] C. d'Allessandro & P. Mertens
Automatic pitch contour stylization using a model of tonal perception
Computer Speech and Language, 9:257--288, 1995.

[auberge:these91] V. Aubergé
La synthèse de la parole : "des règles aux lexiques"
PhD thesis, Université Pierre Mendès-France, Grenoble -France, 1991.

[auberge:tm92] V. Aubergé
Developing a structured lexicon for synthesis of prosody
Talking Machines: Theories, Models and Designs, G. Bailly and C. Benoît, editors, pages 307--321. Elsevier B.V., 1992.

[auberge:lund93] V. Aubergé
Prosody modeling with a dynamic lexicon of intonative forms: Application for text-to-speech synthesis
Working Papers of Lund University, 41:62--66, 1993.

[auberge-etal:eurospeech97] V. Aubergé, T. Crépillat & A. Rilliard
Can we perceive attitudes before the end of sentences? The gating paradigm for prosodic contours
European Conference on Speech Communication and Technology, vol 2, pages 871--874, Rhodes - Greece, 1997.

[badin-etal:jasa90] P. Badin, L.-J. Boë, P. Perrier & C. Abry
Acoustic considerations upon formant convergence
Journal of the Acoustical Society of America, 87(3):1290--1300, 1990.

[bartkova-sorin:speechcom87] K. Bartkova & C. Sorin
A model of segmental duration for speech synthesis in French
Speech Communication, 6:245--260, 1987.

.[beaugendre:these94] F. Beaugendre
Une étude perceptive de l'intonation du Français. Développement d'un modèle et application à la génération automatique de l'intonation pour un système de synthèse à partir du texte
Thèse de troisième cycle, Université Paris XI, Paris - France, 1994.

[beckman-ayers:tobi94] M.E. Beckman & G.M. Ayers
Guidelines to ToBI labelling. Version 2.0

[beckman-hirschberg:tobi94] M.E. Beckman & J. Hirschberg
The ToBI annotation conventions

[beckman-ayers:tobi97] M.E. Beckman & G.M. Ayers
Guidelines to ToBI labelling. Version 3.0
http://www.ling.ohio-state.edu/phonetics/E_ToBI

[beckman:tm97] M.E. Beckman
Speech models and speech synthesis
Progress in Speech Synthesis Jan P. H. van Santen, Richard W. Sproat, Joseph P.

Olive & Julia Hirschberg, editors, pages 477--493. Springer-Verlag, New York, 1997.
[berthier-etal:icphs91] V. Berthier, C. Abry & T.Lallouache.
Coordination du geste et de la parole dans la production d'un instrument traditionnel.
International Congress of Phonetic Sciences, vol 4, pages 34--37, Aix-en-Provence, France, 1991.
[benoit-etal:scom96] C. Benoît, M. Grice & V. Hazan.
The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences.
Speech Communication, 18:381--392, 1996.
[boe-etal:dft94] L.-J. Boë, J._L. Schwartz & N. Vallée
The prediction of vowel systems: perceptual contrast and stability
Fundamentals of speech synthesis and speech recognition, In Eric Keller, editor, pages 185--214. John Wiley and Sons, Chichester, 1994.
[bolinger:89] D. Bolinger
Intonation and its uses
Edward Arnold, London, 1989.
[bolinger:sp81] D. Bolinger
Some intonation stereotypes in English
Studia Phonetica, 18:97--101, 1981.
[bolinger:word51] D. Bolinger
Intonation : Levels versus configuration
Word, VII:199--210, 1951.
[boltz:rpp92] M. Boltz
Temporal expectancies and melody recognition
International Workshop on Rhythm Perception and Production, Ville de Bourges, pages 97--103, 1992.
[bonneau-etal:jasa96] A. Bonneau, L. Djezzar & Y. Laprie
Perception of place of articulation of French stop bursts
Journal of Acoustical Society of America, 100(1):555--564, 1996.
[browman-goldstein:85] C. P. Browman and L. M. Goldstein.
Dynamic modeling of phonetic structure.
Phonetic Linguistics: Essays in honor of Peter Ladefoged, Victoria A. Fromkin, editor, pages 35--53. Academic Press, Inc., 1985.
[browman-goldstein:jphon90] C. P. Browman and L. M. Goldstein.
Gestural specification using dinamically-defined articulatory structures.
Journal of Phonetics, 18(3):299--320, 1990.
[browman-goldstein:py86] C. P. Browman and L. M. Goldstein.
Towards an articulatory phonology.
Phonology Yearbook, 3:219--252, 1986.
[browman-goldstein:phono89] C. P. Browman & L. M. Goldstein.
Articulatory gestures as phonological units.
Phonology, 6:201--251, 1989.
[bull:these97] M. Bull
The timing and coordination of turn-taking
PhD, University of Edimburgh, 1997
[campbell-isard:jphon91] W. N. Campbell & S. D. Isard
Segment durations in a syllable frame
Journal of Phonetics, 19:37--47, 1991.
[campbell:atr96] W. N. Campbell
Synthesizing spontaneous speech

Computing prosody: Computational models for processing spontaneous speech, Yoshinori Sagisaka, Nick Campbell, and Norio Higuchi, editors, pages 165--186. Springer Verlag, 1997.

[campbell:tm92] W. N. Campbell
Syllable-based segmental duration
Talking Machines: Theories, Models and Designs, G. Bailly and C. Benoît, editors, pages 211--224. Elsevier B.V., 1992.

[carlson-granstrom:scom75] R. Carlson & B. Granström
A phonetically oriented programming language for rule description of speech
Speech Communication, pages 245--253, 1975.

[carlson-granstrom:icassp76] R. Carlson & B. Granström
A text-to-speech system based entirely on rules
IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 686--688, 1976.

[changeux-connes:92] J.P. Changeux & A. Connes
Matière à penser
Odile Jacob, 288 pages, ISBN 2020146762.

[coleman:tm92] J. Coleman
« synthesis-by-rule » without segments or rewrite-rules
Talking Machines: Theories, Models and Designs, G. Bailly & C. Benoît, editors, pages 43--60. Elsevier B.V., 1992.

[couper-kuhlen:icphs91] E. Couper-Kuhlen
A rhythm-based metric for turn-taking
International Conference of Phonetic Sciences, vol 1, pages 275--278, Aix-en-Provence, France, 1991.

[damasio:95] A. R. Damasio
L'erreur de Descartes
Editions Odile Jacob, Paris, 1995.

[dauer:jphon83] R. M. Dauer.
Stress-timing and syllable-timing re-analyzed.
Journal of Phonetics, 11:51--62, 1983.

[davis-macneilage:jshr95] B. L. Davis & P. F. MacNeilage
The articulatory basis of babbling
Journal of Speech and Hearing Research, 38:1199--1211, 1995.

[dusterhoff-black:etrw97] K. Dusterhoff & A. Black
Generating F0 contours for speech synthesis using the tilt intonation theory
ETRW Workshop on Prosody, pages 107--110, Athens - Greece, 1997.

[davis-macneilage:ls94] B. L. Davis and P. F. MacNeilage
Organization of babbling: a case study
Language and Speech, 37(4):341--355, 1994.

[dicristo:these85] A. Di Cristo
De la microprosodie à l'intonosyntaxe
Thèse de troisième cycle, Université de Provence, Aix-en-Provence, France, 1985.

[dutoit-etal:icslp96] T. Dutoit, V. Pagel, N. Pierret, F. Bataille & O. van der Vrecken
The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes
International Conference on Speech and Language Processing, vol 2, pages 1393 -- 1396, USA, 1996.

[emerard:these77] F. Emerard
Synthèse par diphones et traitement de la prosodie

Thèse de troisième cycle, Université de Grenoble III,Grenoble, France, 1977.

[emerard-etal:tm92] F. Emerard, L. Mortamet & A. Cozannet
Prosodic processing in a text-to-speech synthesis system using a database and learning procedures
Talking Machines: Theories, Models and Designs, G. Bailly and C. Benoît, editors, pages 225--254. Elsevier B.V., 1992.

[erler-deng:csl93] K. Erler & L. Deng
Hidden Markov model representation of quantized articulatory features for speech recognition
Computer, Speech and Language, 7(3): 265--282, 1993.

[erler-freeman:jasa96] K. Erler & G.H. Freeman
An HMM-based speech recognizer using overlapping articulatory features
Journal of the Acoustical Society of America, 100(4):2500--2513, 1996.

[fant-kruckenberg:kth89] G. Fant & A. Kruckenberg
Preliminaries to the study of Swedish prose reading and reading style
Technical Report, 2, Speech Transmission Laboratory - Department of Speech Communication and Music Acoustics - KTH, Stockholm - Sweden, 1989.

[fant:1960] G. Fant
Acoustic theory of speech production
The Hague: Mouton, 1960.

[fant:icphs91] G. Fant
Units of temporal organization. stress groups versus syllables and words
International Congress of Phonetic Sciences, vol 1, pages 247--250, Aix-en-Provence, France, 1991.

[fant-etal:jphon91] G. Fant, A. Kruckenberg & L. Nord
Durational correlates of stress in Swedish, French and English
Journal of Phonetics, 19:351--361, 1991.

[fant:icslp92] G. Fant
Vocal tract area functions of Swedish vowels and a new three-parameter model.
International Conference on Speech and Language Processing, vol 1, pages 807--810, Edmonton - Alberta, 1992.

[fant-kruckenberg:icslp96] G. Fant & A. Kruckenberg
On the quantal nature of speech timing
International Conference on Speech and Language Processing, vol 3, pages 2044--2047, Philadelphia - USA, 1996.

[fonagy:82] I. Fónagy
Situation et signification. Prolégomènes à un dictionnaire des énoncés en situation
Benjamins, Amsterdam, 1982.

[fonagy-etal:fl84] I. Fónagy, E. Bérard & J.Fónagy
Clichés mélodiques
Folia Linguistica, 17:153--185, 1984.

[fonagy:icphs87] I. Fónagy
Semantic diversity in intonation
International Congress of Phonetic Sciences, vol 2, pages 468--471, Tallinn, Estonia-USSR, 1987.

.[fowler:haskins90] C. A. Fowler
Listener-talker attunements in speech
Haskins Laboratories Staturs Report on Speech research,1:11--129, 1990.

.[fowler:jphon83] C. A. Fowler
Realism and unrealism: a reply

Journal of Phonetics, 11(4):303--322, 1983.

[fowler:jasa91] C. A. Fowler

Auditory perception is not special: We see the world, we fell the world, we hear the world

Journal of the Acoustical Society of America, 89(6):2910--2915, 1991.

[fowler:jasa96] C.A. Fowler

Listeners do hear sounds, not tongues

Journal of the Acoustical Society of America, 99(3):1730--1741, 1996.

[fraisse:74] P. Fraisse

La psychologie du rythme

Presses Universitaires de France, Paris, 1974.

[fraisse:80] P. Fraisse

Des synchronisations sensori-motrices aux rythmes

Anticipation et Comportement, J. Requin, editor, pages 233--257. Editions du CNRS, Paris, 1980.

[fujisaki-sudo:areri71] H. Fujisaki & H. Sudo

A generative model for the prosody of connected speech in Japanese

Annual Report of Engineering Research Institute, 30:75--80, 1971.

[fujisaki-kawai:icassp88] H. Fujisaki & H. Kawai

Realization of linguistic information in the voice fundamental frequency contour of the spoken japanese

IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 663--666, 1988.

[gee-grosjean:cp83] J.-P. Gee & F. Grosjean

Performance structures: a psycholinguistic and linguistic appraisal

Cognitive Psychology, 15:411--458, 1983.

[gerken-etal:cog94] L. A. Gerken, P. W. Jusczyk, and D. R. Mandel

When prosody fails to cue syntactic structure: 9-months-olds. Sensitivity to phonological versus syntactic phrases

Cognition, 51:237--265, 1994.

[gronnum:92] N. Grönnum

The ground-works of Danish intonation

Museum Tusculanum Press - Univ. Copenhagen, Copenhagen, 1992.

[grosjean-lane:cp79] F. Grosjean & X. Lane

The patterns of silence: Performance structures in sentence production.

Cognitive Psychology, 11:58--81, 1979.

[grosjean:ling83] F. Grosjean

How long is the sentence? prediction and prosody in the on-line processing of language

Linguistica, 21:501--529, 1983.

[grosjean-hirt:lcp96] F. Grosjean & C. Hirt

Using prosody to predict the end of sentences in English and French: normal and brain-damaged subjects

Language and Cognitive Processes, 11(1):107--134, 1996

[guenther:icphs95] F. H. Guenther

A modeling framework for speech motor development and kinematic articulator control

International Congress of Phonetic Sciences, vol 2, pages 93--99, Stockholm - Sweden, 1995.

[guenther:pr95] F.H. Guenther

Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production
Psychological Review, 102(3):594--621, 1995.

[guenther:these92] F. H. Guenther
Neural models of adaptive sensori-motor control for flexible reaching and speaking
PhD thesis, Boston University - Boston, 1992.

[hary-moore:biocyber87] D. Hary & G. P. Moore
Synchronizing human movement with an external clock source
Biological Cybernetics, 56:305--311, 1987.

[hermansky-broad:icassp92] H. Hermansky & D. J. Broad
The effective second formant f'2 and the vocal tract front-cavity
IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 480--483, 1989.

[hermansky-etal:icassp95] H. Hermansky, E. Wan & C. Avendano
Speech enhancement based on temporal processing
IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 405—408, 1995.

[hertz-etal:assp85] S. Hertz, J. Kadin & K. Karplus
The delta rule development system for speech synthesis from text
IEEE Transactions on Acoustics, Speech and Signal Processing, 73(11):1589--1601, 1985.

[vanheuven:etrw97] V. J. van Heuven, J. Haan, E. Janse & E. J. van der Torre
Perceptual identification of sentence type and the time-distribution of prosodic interrogativity marker in Dutch
ETRW Workshop on Prosody, pages 317--320, Athens - Greece, 1997.

[hirst:icphs91] D. Hirst
Intonation models: towards a third generation
International Congress of Phonetic Sciences, vol 1, pages 305--310, Aix-en-Provence, France, 1991.

[hirst:lund93] D. Hirst
Peak, boundary and cohesion characteristics of prosodic grouping
Working Papers of Lund University, 41:32--37, 1993.

[hirst:cutler-ladd83] D. Hirst
Structures and categories in prosodic representations
Prosody: models and measurements, A. Cutler and A. Ladd, editors, , pages 93--109. Springer-Verlag, Berlin, 1983.

[jones-boltz:psyr89] M. R. Jones & M. G. Boltz
Dynamic attending and responses to time
Psychological Review, 96:459--491, 1989.

[jordan:coins88] M. I. Jordan
Supervised learning and systems with excess degrees of freedom
COINS Tech. Rep. 88-27, University of Massachussetts, Computer and Information Sciences, Amherst - MA - USA, 1988.

[jordan-rumelhart:91] M. I. Jordan & D. E. Rumelhart
Forward models: Supervised learning with a distal teacher
Occasional Paper 40, MIT, Center for Cognitive Sciences, Cambridge - USA, 1991.

[jordan:attention90] M. I. Jordan
Motor learning and the degrees of freedom problem
Attention and Performance, Marc Jeannerod, editor, , vol XIII. Lawrence Erlbaum Associates, Hillsdale, NJ - USA, 1990.

[klatt:jasa87] D. H. Klatt
Review of text-to-speech conversion for English
Journal of the Acoustical Society of America, 82(3):737--793, 1987.
[kluender:perilus91] K. R. Kluender.
Psychoacoustic complementarity and the dynamics of speech perception and production.
PERILUS XIV - Publication of the Department of Linguistics, pages 131--135, 1991.
[konopczynski:icphs91] G. Konopczynski
Acquisition de la proéminence dans le langage émergent
International Congress of Phonetic Sciences, vol 1, pages 333--337, Aix-en-Provence, France, August 1991.
[kuhn:jasa75] G. M. Kuhn
On the front-cavity resonance and its possible role in speech perception
Journal of the Acoustical Society of America, 68:578--585, 1975.
[kuhn:jasa79] G. M. Kuhn
Stop consonant place perception with single-formant stimuli: Evidence for the role of the front-cavity resonance
Journal of the Acoustical Society of America, 65(3):774--788, 1979.
[kewley-port-etal:jasa83] D. Kewley-Port, D. B. Pisoni & M. Studdert-Kennedy
Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants
Journal of the Acoustical Society of America, 73(5):1779--1793, 1983.
[ladd:jasa88] D. R. Ladd
Declination « reset » and the hierarchical organization of utterances
Journal of the Acoustical Society of America, 84(2):530--544, 1988.
[lin-fant:eurospeech89] Q. Lin & G. Fant
Vocal tract area function parameters from formant frequencies
European Conference on Speech Communication and Technology, 2:673--676, 1989.
[lindblom-lindgren:perilus85] B. Lindblom & R. Lindgren
Speaker-listener interaction and phonetic variation
PERILUS IV - Publication of the Department of Linguistics, 1985.
[lindblom:icphs87] B. Lindblom
Adaptive variability and absolute constancy in speech signals: two themes in the quest for phonetic invariance
International Congress of Phonetic Sciences, vol 3, pages 9--18, Tallin, Estonia, 1987
[liljencrants-lindblom:language72] J. Liljencrants and B. Lindblom
Numerical simulation of vowel quality systems: The role of perceptual contrasts
Language, 48:839--861, 1972.
[local:94] J. Local
Phonological structure, parametric phonetic interpretation and natural-sounding synthesis
Fundamentals of speech synthesis and speech recognition, Eric Keller Editor, 25--270, 1994.
[marcus:these76] S. M. Marcus.
Perceptual centres
PhD thesis, Cambridge University, 1976.
[marcus:pp81] S. M. Marcus.
Acoustic determinants of perceptual center (p-center) location.
Perception and Psychophysics, 30(3):247--256, 1981.
[markey-etal:ld95] K. L. Markey, L. Menn & M. C. Mozer

A developmental model of the sensorimotor foundations of phonology
Proceedings of the 19th Boston University Conference on Language Development, D. MacLaughlin and S. McEwen, editors, , pages 367--378, Cascadilla Press, Somerville - MA, 1995.

[markey:cs94] K. L. Markey
Acoustic-based syllabic representation and articulatory gesture detection: prerequisites for early childhood phonetic and articulatory development
Proceedings of the 16th Annual Conference of the Cognitive Science Society In A. Ram and K. Eiselt, editors, , pages 595--600. Lawrence Erlbaum Associates, Hillsdale, NJ - USA, 1994.

[markey:these94] K. L. Markey
The sensorimotor foundations of phonology: a computational model of early childhood articulatory and phonetic development
PhD thesis, University of Colorado - Boulder, Colorado - USA, 1994.

[marsi-etal:tm97] E. C. Marsi, P.-A. J. M. Coppen, C. H. M. Gussenhoven & T. C. M. Rietveld
Prosodic and intonational domains in speech synthesis
Progress in Speech Synthesis Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive & Julia Hirschberg, editors, pages 477--493. Springer-Verlag, New York, 1997.

[mcgowan:scom94] R. McGowan
Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: preliminary model tests
Speech Communication, 14:19--48, 1994.

[merleau-ponty:45] M. Merleau-Ponty
Phénoménologie de la perception
Gallimard, Paris, 1945.

[morasso-sangui:ms97] P. Morasso & V. Sanguineti
From cortical maps to the control of muscles
Self-Organization, Computational Maps and Motor Control, Pietro Morasso & Vittorio Sanguineti, editors, pages 547--591, Elsevier, Amsterdam, 1997.

[morgan-demuth:96] J. L. Morgan and K. Demuth
Signal to syntax: an overview
Lawrence Erlbaum Associates, Mahwah, NJ - USA, 1996.

[morton-etal:pr76] J. Morton, S. Marcus & C. Frankish
Perceptual centers (p-centers)
Psychological Revue, 83(5):405--408, 1976.

[monnin-grosjean:ap93] P. Monnin & F. Grosjean.
Les structures de performance en français : caractérisation et prédiction.
L'Année Psychologique, 93:9--30, 1993.

[mussa-ivaldi-bizzi:ms97] F. A. Mussa-Ivaldi & E. Bizzi
Learning Newtonian mechanics
Self-Organization, Computational Maps and Motor Control, Pietro Morasso & Vittorio Sanguineti, editors, pages 191--237, Elsevier, Amsterdam, 1997.

[nossair-zahorian:jasa91] Z. B. Nossair & S. A. Zahorian.
Dynamic spectral shape features as acoustic correlates for initial stop consonants.
Journal of the Acoustical Society of America, 89:2978--2991, 1991.

[ohman:jasa67] S. E. G. Öhman
Numerical model of coarticulation
Journal of the Acoustical Society of America, 41:310--320,1967.

[olaszy-etal:tm92] G. Olaszy, G. Gordos, and G. Németh

The Multivox multilingual text-to-speech converter
Talking Machines: Theories, Models and Designs, G. Bailly and C. Benoît, editors, pages 385--411. Elsevier B.V., 1992.

[oshaughnessy:jphon81] D. O'Shaughnessy
A study of French vowel and consonant durations
Journal of Phonetics, 9:385--406, 1981.

[oshaughnessy:jasa84] D. O'Shaughnessy
A multispeaker analysis of durations in read French paragraphs
Journal of the Acoustical Society of America, 76(6):1664--1672, 1984.

[perrier-etal:jshr96] P. Perrier, D. J. Ostry & R. Laboissière.
The Equilibrium Point Hypothesis and its application to speech motor control.
Journal of Speech and Hearing Reserach, 39:365--377, 1996.

[petitot:85] J. Petitot
Les catastrophes de la parole. De Roman Jakobson à René Thom
Maloine, Paris, 1985.

[petitot:86] J. Petitot
Le « morphological turn » de la phénoménologie
Morphogénèse du sens, II, Centre d'analyse et de Mathématique Sociales, EHESS-CNRS, 1986.

[petitot:rs90] J.Petitot
Le physique, le morphologique, le symbolique - remarques sur la vision
Revue de Synthèse, IV(1-2):139--183, 1990.

[pfitzinger:icslp98] H.R. Pfitzinger
Local speech rate as a combinaison of syllable and phone rate
International Conference on Speech and Language Processing, 3, 1087--1090.

[pierrehumbert:jasa81] J. Pierrehumbert.
Synthetizing intonation.
Journal of the Acoustical Society of America, 70(4):985--995, 1981.

[pierrehumbert:jphon90] J. Pierrehumbert
Phonological and phonetic representation
Journal of Phonetics, 19:375--394, 1990.

[pompino-marschall:jphon89] B. Pompino-Marschall
On the psychoacoustic nature of the p-center phenomenon
Journal of Phonetics, 17:175--192, 1989.

[rossi:speechcom93] M. Rossi.
A model for predicting the prosody of spontaneous speech (ppss model).
Speech Communication, 13:87--107, 1993.

[riley:tm92] M. Riley
Tree-based modelling of segmental durations
Talking Machines: Theories, Models and Designs, G. Bailly and C. Benoît, editors, pages 265--274. Elsevier B.V., 1992.

[saerens-etal:icphs91] M. Saerens, A. Soquet & P. Jospa
Trying to determine place of articulation of plosives with a vocal tract model
International Congress of Phonetic Sciences, 2:66--69, 1991.

[schwartz-escudier:sc89] J.-L. Schwartz & P. Escudier
A strong evidence for the existence of a large-scale integrated spectral representation in vowel perception
Speech Communication, 8:235--259, 1989

[schwartz-etal:jphon93] J.-L. Schwartz, D. Beautemps, C. Abry & P. Escudier
Inter-individual and cross-linguistic strategies for the production of the [i] vs [y]

contrast
Journal of Phonetics, 21:411--425, 1993.

[scott:these93] S. Scott
Perceptual centres in speech - an acoustic analysis
PhD thesis, University College, London, England, 1993.

[silverman-etal:icslp92] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert & J.Hirschberg
ToBI: a standard for labeling english prosody
International Conference on Speech and Language Processing, 2:867--870, 1992.

[smith:jphon78] B.L. Smith.
Temporal aspects of English speech production: a developmental perspective.
Journal of Phonetics, 1(6):37--67, 1978.

[smits:icslp96] R. Smits.
Context-dependent relevance of burst and transitions for perceived place in stops: it's in production, not perception.
International Conference on Speech and Language Processing, pages 2470--2473, Philadelphia - USA, 1996.

[smits:thesis95] R. Smits
Detailed versus gross spectro-temporal cues for the perception of stop consonants.
PhD thesis, Institute of Perception Research (IPO), Eindhoven - Netherlands, 1995.

[smits:jphon95] R. Smits & L. ten Bosch.
A note on classification experiments in acoustic phonetics.
Journal of Phonetics, 23:477--485, 1995.

[smits-etal:jasa96a] R. Smits, L. ten Bosh & R. Collier
Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. perception experiment.
Journal of the Acoustical Society of America, 100(6):3852--3864, 1996.

[smits-etal:jasa96b] R. Smits, L. ten Bosh & R. Collier.
Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. II. modeling and evaluation.
Journal of the Acoustical Society of America, 100(6):3865--3881, 1996.

[smolensky92] P. Smolensky
IA connexionniste, IA symbolique et cerveau
Introduction aux sciences cognitives, D. Andler, Editor, Gallimard, Paris, pages 77-106, 1992.

[sorokin:scom92] V. N. Sorokin
Determination of vocal tract shape for vowels
Speech Communication, 11:71--85, 1992.

[sorokin:scom94] V. N. Sorokin
Inverse problems for fricatives
Speech Communication, 14:249--262, 1994.

[stevens:72] K. N. Stevens
The quantal nature of speech: Evidence from articulatory-acoustic data
Human Communication: A unified view, E. E. David Jr. and P. B. Denes, editors, , pages 51--66. McGraw-Hill, New York, 1972

[stevens:perilus91] K. N. Stevens
Speech perception based on acoustic landmarks: Implications for speech production
PERILUS XIV - Publication of the Department of Linguistics, pages 83--88, 1991.

[stevens-bickley:jphon91] K. N. Stevens & C. A. Bickley
Constraints among parameters simplify control of Klatt formant synthesizer

Journal of Phonetics, 19:161--174, 1991.

[sussman:phonetica91] H. M. Sussman
The representation of stop consonants in three-dimensional acoustic space
Phonetica, 48(1):18--31, 1991.

[sussman-etal:jasa97] H. M. Sussman, N. Bessell, E. Dalston & T.Majors
An investigation of stop place of articulation as a function of syllable position : a locus equations perspective
Journal of the Acoustical Society of America,101(5):2826--2838, 1997.

[sussman-etal:jasa93] H. M. Sussman, K. A. Hoemeke & F. S. Ahmed
A cross-liguistic investigation of locus equations as a phonetic descriptor for place of articulation
Journal of the Acoustical Society of America, 94(3):1256--1268,1993.

[sussman-etal:jasa91] H. M. Sussman, H. A. McCaffrey & S. A. Matthews
An investigation of locus equations as a source of relational invariance for stop place categorization
Journal of the Acoustical Society of America, 90(3):1309--1325, 1991.

[taylor-etal:etrw98] P. Taylor, A. Black & R. Caley
The architecture of the Festival speech synthesis system
3rd ESCA/COCOSDA International Workshop on Speech Synthesis, pages 249--253, Jenolan Caves, Australia, 1998.

[taylor-black:etrw94] P. Taylor & A. Black
Synthesizing conversational intonation from a linguistically rich input
2nd ESCA/COCOSDA International Workshop on Speech Synthesis, pages 131--134, New Paltz - USA, 1994.

[thorsen:fl83] N. G. Thorsen.
Standard Danish sentence intonation - Phonetic data and their representation.
Folia Linguistica, 17:187--220, 1983.

[turvey-etal:90] M.T. Turvey, R.C. Schmidt & L. Rosenblum
Clock and motor components in absolute coordination of rhythmic movements
Haskins Laboratories Status Report on Speech Research, pages 231--242, 1990.

[vancoile:icassp89] B. van Coile
The depes development system for text-to-speech synthesis
IEEE International Conference on Acoustics, Speech, and Signal Processing, vol 4, pages 250--253, 1989.

[vansanten:tm92] J. P. H. Van Santen
Deriving text-to-speech durations from natural speech
Talking Machines: Theories, Models and Designs, G. Bailly and C. Benoît, editors, pages 275--285. Elsevier B.V., 1992.

[vansanten:sc92] J. P. H. Van Santen
Contextual effects on vowel duration
Speech Communication, 11:513--546, 1992.

[vansanten:csl94] J. P. H. Van Santen
Assignment of segmental duration in text-to-speech synthesis
Computer Speech and Language, 8:95--128, 1994.

[vansanten:atr96] J. P. H. Van Santen
Segmental duration and speech timing
Computing prosody: Computational models for processing spontaneous speech, Y. Sagisaka, N. Campbell & N. Higuchi, editors, pages 225--249. Springer Verlag, 1997.

[villiers:93] J.-M.-M.-P.-A. de Villiers de l'Isle-Adam
L'Eve future

Réédité en 1993 par Gallimard, Paris, 1886.
[viviani-stucchi:advpsy92] P. Viviani & N. Stucchi
Motor-perceptual interactions
Advances in psychology : tutorials in motor behavior II, G.E. Stelmach & J. Requin, editors, pages 222—248, North-Holland, Amsterdam, 1992.
[zahorian-jagharghi:jasa91] St. A. Zahorian & A. J. Jagharghi
Speaker normalization of static and dynamic vowel spectral featuress
Journal of the Acoustical Society of America, 90:67--75, 1991.

BIBLIOGRAPHIE PERSONNELLE

[bailly:these] Gérard Bailly
Contribution à la détermination automatique de la prosodie du français parlé à partir d'une analyse syntaxique. Etablissement d'un modèle de génération
Thèse de troisième cycle, Institut National Polytechnique de Grenoble, Grenoble, France, 1983.

[bailly:jep86a] Gérard Bailly
Détection du fondamental par amdf et programmation dynamique
Journées d'Etudes sur la Parole, pages 285--288, Aix-en-Provence - France, 1986. GALF.

[bailly:jep86b] Gérard Bailly
Un modèle de congruence relationnel pour la synthèse de la prosodie du français
Journées d'Etudes sur la Parole, pages 75--78, Aix-en-Provence - France, 1986. GALF.

[bailly-liu:jep87] Gérard Bailly & Janping Liu
Détection d'indices par quantification vectorielle et réseaux Markoviens
Journées d'Etudes sur la Parole, pages 60--63, Hammamet- Tunisie, 1987. GALF.

[aldakkak-etal:iasted87] Omaima Al Dakkak, Gérardo Murillo & Gérard Bailly
Automatic extraction of formant parameters using a-priori knowledge
IASTED, Applied Control Filtering and Signal Processing, Geneva - Switzerland, 1987

[aldakkak-etal:nato87] Omaima Al Dakkak, Gérardo Murillo, Gérard Bailly & Bernard Guérin
Using contextual information in view of formant analysis improvement
Recent advances in Speech Understanding and dialog systems, NATO ASI Series, Bad Windsheim - Germany, 1987.

[bailly-etal:fase88] Gérard Bailly, Gérardo Murillo, Omaima Al Dakkak &Bernard Guérin
A text-to-speech synthesis system for French by formant synthesis
7th FASE Symposium, pages 225--260, 1988.

[bailly-etal:icp88] Gérard Bailly, Adeline Perrin & Yves Lepage
Common approaches in speech synthesis and automatic translation of text
Bulletin du Laboratoire de la Communication Parlée - Grenoble , 2B:295--311, 1988.

[aldakkak-etal:icp88] Omaima. Al Dakkak, Gérardo Murillo, Gérard Bailly & Bernard Guérin
A database of formant parameters for knowledge extraction and synthesis-by-rule
Bulletin du Laboratoire de la Communication Parlée , pages 391--405, 1988.

[marteau-etal:icassp88] Pierre-François Marteau, Gérard Bailly & Marie-Thérèse Janot-Giorgetti
Stochastic model of diphone-like segments based on trajectory concepts
IEEE International Conference on Acoustics, Speech & Signal Processing, vol 6, pages 615--618, New York - USA, 1988.

[bailly:calliope89] Gérard Bailly
Synthèse de la parole
La parole et son traitement automatique, In Jean-Pierre Tubach, editor, pages 408--448. Masson, Paris, 1989.

[bailly:speechcom89] Gérard Bailly

Integration of rhythmic and syntactic constraints in a model of generation of French prosody
Speech Communication , 8:137--146, 1989.

[bailly-liu:greco89] Gérard Bailly & Janping Liu
Détection de formants par quantification vectorielle et réseaux markoviens
Actes du séminaire Décodage Acoustico-Phonétique , Greco Dialogue Homme-Machine, pages 89--94, Nancy - France, 1989.

[bailly-etal:icassp89] Gérard Bailly, Pierre-François Marteau & Christian Abry
A new algorithm for temporal decomposition of speech. application to a numerical model of coarticulation
IEEE International Conference on Acoustics, Speech & Signal Processing , pages 508--511, 1989.

[bailly-tran:eurospeech89] Gérard Bailly & Alain Tran
Compost: a rule-compiler for speech synthesis
European Conference on Speech Communication and Technology , vol 1, pages 136--139, 1989.

[bailly-etal:nsi90] Gérard Bailly, Morton Bach, Morton Olesen, Jean-Luc Schwartz & Andrew Morris
Génération de trajectoires articulatoires par réseau séquentiel
5èmes Journées NSI , pages 191--192, Aussois - France, 1990.

[bailly-etal:asa90a] Gérard Bailly, Thierry Barbe, Stéphane Veste, Haïdong Wang & Denis Tuffelli
Automatic labelling of large databases: tools and methodology
Meeting of the Acoustical Society of America, 87:S, 1990.

[bailly-etal:asa90b] Gérard Bailly, Michael Jordan, Marios Mantakas, Jean-Luc Schwartz, Morton Bach & Morton Olesen
Simulation of vocalic gestures using an articulatory model driven by a sequentiel neural network
Meeting of the Acoustical Society of America, 87:S105, 1990.

[bailly-etal:etrw90] Gérard Bailly, Thierry Barbe & Haïdong Wang
Automatic labelling of large prosodic databases: tools, methodology and links with a text-to-speech system
ETRW Workshop on Speech Synthesis , pages 201--204, 1990.

[bailly-guerti:jep90] Gérard Bailly & Mhania Guerti
Anticipation et rétention dans les mouvements vocaliques en Français
Journées d'Etudes sur la Parole, pages 292--295, Montréal - Canada, 1990.

[bailly-liu:ja90] Gérard Bailly & Janping Liu
Détection d'indices par quantification vectorielle et réseaux Markoviens
Journal d'Acoustique, 3:143--151, 1990.

[barbe-bailly:jep90] Thierry Barbe & Gérard Bailly
Evaluation d'un détecteur de fréquence fondamentale du signal microphonique par comparaison à une référence laryngographique
Journées d'Etudes sur la Parole, pages 165--169, Montréal, 1990.

[wang-etal:asa90] Haïdong Wang, Gérard Bailly & Denis Tuffelli
Automatic segmentation and alignment of continuous speech based on the temporal decomposition model
Meeting of the Acoustical Society of America, 87:S106, 1990.

[wang-etal:icassp90] Haïdong Wang, Gérard Bailly & Denis Tuffelli

Automatic segmentation and alignment of continuous speech based on the temporal decomposition model
International Conference on Speech and Language Processing , vol 1, pages 457--460, Kobe - Japan, 1990.

[ye-etal:aann90] Hayan Yé, Shenrui Wang, Gérard Bailly & François Robert
Exploration of temporal processing of a sequential network for speech parameter estimation
Proceedings of Applications of Artificial Neural Networks , pages 16--20, Orlando, Florida, 1990.

[alissali-bailly:rfia91] Mamoun Alissali & Gérard Bailly
Compost: un serveur de synthèse multilingue
8e Congrès sur la Reconnaissance de Formes et l'Intelligence Artificielle , pages 183--192, Lyon-Villeurbanne, 1991.

[bailly-guerti:icphs91] Gérard Bailly & Mhania Guerti
Synthesis-by-rule for French
12th International Congress of Phonetic Sciences , pages 506--509, Aix-en-Provence, France, August 1991.

[bailly-etal:jphon91] Gérard Bailly, Rafaël Laboissière & Jean-Luc Schwartz
Formant trajectories as audible gestures: An alternative for speech synthesis
Journal of Phonetics , 19(1):9--23, 1991.

[guerti-bailly:eurospeech91] Mhania Guerti & Gérard Bailly
Synthesis-by-rule using compost: modelling resonance trajectories
European Conference on Speech Communication and Technology, pages 43--46, Genova - Italy, 1991.

[laboissiere-etal:cmss91] Rafaël Laboissière, Jean-Luc Schwartz & Gérard Bailly
Motor control for speech skills: A connectionist approach
Proceedings of the 1990 Connectionist Models Summer School In David S. Touretzky, Jeffrey L. Elman, Terrence J. Sejnowski & Geoffrey E. Hinton, editors, pages 319--327. Morgan Kaufmann, San Mateo, CA, 1991.

[laboissiere-etal:perilus91] Rafaël Laboissière, Jean-Luc Schwartz & Gérard Bailly
Modelling the speaker-listener interaction in a quantitative model for speech motor control: a framework and some preliminary results
PERILUS XIV, Department of Linguistics, pages 57--62, 1991.

[bailly-etal:eusipco92] Gérard Bailly, Christian Abry, Louis-Jean Boë , Rafaël Laboissière, Pascal Perrier & Jean-Luc Schwartz
Inversion and speech recognition
Signal Processing VI: Theories and Applications, J. Vandewalle, R. Boîte, M. Moonen & A. Oosterlinck, editors, vol 1, pages 159--164. Elsevier, Amsterdam, 1992.

[bailly-alissali:ts92] Gérard Bailly & Mamoun Alissali
Compost: a server for multilingual text-to-speech system
Traitement du Signal, 9(4):359--366, 1992.

[bailly-etal:tm92] Gérard Bailly, Thierry Barbe & Haïdong Wang
Automatic labelling of large prosodic databases: tools, methodology and links with a text-to-speech system
Talking Machines: Theories, Models and Designs, Gérard Bailly and Christian Benoît, editors, pages 323--333. Elsevier B.V., 1992.

[bailly-benoit:tm92] Gérard Bailly & Christian Benoît
Talking Machines: Theories, Models and Designs,. Elsevier B.V., 1992.

[barbosa-bailly:bourges] Plínio Barbosa and Gérard Bailly
Generating segmental duration by p-centers
Fourth Rhythm Workshop: Rhythm Perception and Production, Ville de Bourges, Catherine Auxiette, Carolyn Drake & Claire Gérard, editors, pages 163--168, Bourges, France, 1992.
[barbosa-bailly:jep92] Plínio Barbosa & Gérard Bailly
Génération automatique des p-centers
Journées d'Etudes sur la Parole - Bruxelles , pages 357--361, 1992.
[boe-etal:jphon92] Louis-Jean Boë , Pascal Perrier & Gérard Bailly
The geometric vocal tract variables controlled for vowel production: Proposals for constraining acoustic-to-articulatory inversion
Journal of Phonetics , 20(1):27--38, 1992
[laboissiere-etal:arc92] Rafaël Laboissière, Jean-Luc Schwartz & Gérard Bailly
Modélisation du contrôle moteur en production de la parole: vers un « robot parlant »
Cinquième Colloque de l'ARC , pages 115--124, Nancy - France, 1992.

[alissali-bailly:eurospeech93] M. Alissali and G. Bailly
Compost: A client-server model for applications using text-to-speech
European Conference on Speech Communication and Technology , vol 3, pages 2095--2098, Berlin- Germany, 1993.
[bailly:eurospeech93] Gérard Bailly
Resonances as possible representation of speech in the auditory-to-articulatory transform
European Conference on Speech Communication and Technology , vol 2, pages 1511--1514, 1993.
[bonnyman-etal:tas93] James Bonnyman, K. M. Curtis & Gérard Bailly
A transputer-based recurrent neural network for resonance tracking of speech
Transputer Applications and Systems , 1:1219--1228, 1993.
[isis93] Emmanuel Reynier & Gérard Bailly
Isis: an interactive system for image and speech on Unix
Technical Report ISIS/1, Institut de la Communication Parlée, Grenoble, France, 1993.

[bailly-etal:etrw94] Gérard Bailly, Eric Castelli & Bernard Gabioud
Building prototypes for articulatory speech synthesis
Second ESCA Workshop on Speech Synthesis , pages 9--12, New Paltz - USA, 1994.
[barbosa-bailly:etrw94] Plínio Barbosa & Gérard Bailly
Generating pauses within the z-score model
Second ESCA Workshop on Speech Synthesis, pages 101--104, New Paltz -USA, 1994.
[barbosa-bailly:scom94] Plínio Barbosa & Gérard Bailly
Characterisation of rhythmic patterns for text-to-speech synthesis
Speech Communication, 15:127--137, 1994.
[bonnyman-etal:icsipnn94] James Bonnyman, K. M. Curtis & Gérard Bailly
A neural network application for the analysis and synthesis of multilingual speech
IEEE International Conference on Speech, Image Processing and Neural Networks , vol 1, pages 327--330, Hong Kong, 1994.

[auberge-bailly:eurospeech95] Véronique Aubergé & Gérard Bailly
Generation of intonation: a global approach
European Conference on Speech Communication and Technology , pages 2065--2068, Madrid, 1995.

[badin-etal:ica95] Pierre Badin, Bernard Gabioud, Denis Beautemps, Tahar Lallouache, Gérard Bailly, Shinji Maeda, Jean-Pierre Zerling & Gilbert Brock
Cineradiography of VCV sequences: articulatory-acoustic data for a speech production model
International Congress on Acoustics , pages 349--352, Trondheim - Norway, 1995.

[bailly:levels95] Gérard Bailly
Caracterisation of formant trajectories by tracking vocal tract resonances
Levels in speech communication : relations and interactions, Christel Sorin, Joseph Mariani, Henri Méloni & Jean Schoentgen, editors, pages 91--102. Elsevier, Amsterdam, 1995.

[bailly:lumigny95] Gérard Bailly
Pistes de recherches en synthèse de la parole
Ecole Thématique : fondements et perspectives en traitement automatique de la parole, Henri Méloni, editor, pages 211--220. Université d'Avignon et des Pays de Vaucluse, Marseille -Luminy - France, 1995.

[bailly:icphs95] Gérard Bailly
Recovering place of articulation for occlusives in VCVs
International Congress of Phonetic Sciences , vol 2, pages 230--233, Stockholm - Sweden, 1995.

[bailly-etal:eurospeech95] Gérard Bailly, Louis-Jean Boë , Nathalie Vallée & Pierre Badin
Articulatori-acoustic prototypes for speech production
European Conference on Speech Communication and Technology , vol 2, pages 1913--1916, Madrid - Spain, 1995.

[morlec-etal: icphs95] Yann Morlec, Véronique Aubergé & Gérard Bailly
Evaluation of automatic generation of prosody with a superposition model
International Congress of Phonetic Sciences, vol 4, pages 224--227, Stockholm - Sweden, 1995.

[morlec-etal:eurospeech95] Yann Morlec, Gérard Bailly & Véronique Aubergé
Synthesis and evaluation of intonation with a superposition model
Proceedings of the European Conference on Speech Communication and Technology, vol 3, pages 2043--2046, Madrid - Spain, 1995.

[badin-etal:etrw96] Pierre Badin, Khaled Mawass, Gérard Bailly, Christophe Vescovi, Denis Beautemps & Xavier Pelorson
Articulatory synthesis of fricative consonants : data and models
ETRW on Speech Production Modelling: from Control Strategies to acoustics , pages 221--224, Autrans - France, 1996.

[bailly:etrw96] Gérard Bailly
Sensori-motor control of speech movements
ETRW on Speech Production Modelling: from Control Strategies to acoustics , pages 145--154, Autrans - France, 1996.

[bailly:aupelf96] Gérard Bailly
Pistes de recherches en synthèse de la parole
Fondements et perspectives en traitement automatique de la parole, Henri Méloni, editor, pages 109--122. AUPELF-UREF, Paris - France, 1996.

[bailly:icslp96] Gérard Bailly
Building sensori-motor prototypes from audio-visual exemplars
International Conference on Speech and Language Processing , pages 957--960, Philadelphia - USA, 1996.

[bailly:jep96] Gérard Bailly

Emergence de prototypes sensori-moteurs à partir d'exemplaires audio-visuels
Journées d'Etudes sur la Parole, pages 87--90, Avignon - France, 1996.

[beautemps-etal:etrw96] Denis Beautemps, Pierre Badin, Gérard Bailly, Arturo Galvàn & Rafaël Laboissière
Evaluation of an articulatory-acoustic model based on a reference subject
ETRW on Speech Production: from Control Strategies to acoustics , pages 45--48, Autrans - France, 1996.

[morlec-etal:icslp96] Yann Morlec, Gérard Bailly & Véronique Aubergé
Generating intonation by superposing gestures
International Conference on Speech and Language Processing , vol 1, pages 283--286, Philadelphia - USA, 1996.

[morlec-etal:jep96] Yann Morlec, Gérard Bailly & Véronique Aubergé
Un modèle connexionniste modulaire pour l'apprentissage des gestes intonatifs
Journées d'Etudes sur la Parole, pages 207--210, Avignon - France, 1996.

[neagu-bailly:jep96] Adrien Neagu and Gérard Bailly
R1, R2 et R3 : un ensemble robuste de paramètres pour la caractérisation des espaces vocaliques
Journées d'Etudes sur la Parole, pages 247--250, Avignon - France, 1996.

[bailly:atr97] Gérard Bailly
No future for comprehensive models of intonation?
Computing prosody: Computational models for processing spontaneous speech, Yoshinori Sagisaka, Nick Campbell & Norio Higuchi, editors, pages 157--164. Springer Verlag, 1997.

[bailly-auberge:tm97] Gérard Bailly & Véronique Aubergé
Phonetic and phonological representations for intonation
Progress in Speech Synthesis, Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive & Julia Hirschberg, editors, pages 435--441. Springer Verlag, New York, 1997.

[bailly-etal:fonagy97] Gérard Bailly, Véronique Aubergé & Yann Morlec
Des représentations cognitives aux représentations phonétiques de l'intonation
Polyphonie pour Iván Fónagy, Jean Perrot, editor, pages 19--28. L'Harmattan, 1997.

[bailly-etal:morasso97] Gérard Bailly, Rafaël Laboissière & Arturo Galván
Learning to speak: Speech production and sensori-motor representations
Self-Organization, Computational Maps and Motor Control, Pietro Morasso & Vittorio Sanguineti, editors, pages 593--615, Elsevier, Amsterdam, 1997.

[barbosa-bailly:tm97] Plínio Barbosa & Gérard Bailly
Generation of pauses within the z-score model
Progress in Speech Synthesis, Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive & Julia Hirschberg, editors, pages 365--381. Springer Verlag, New York, 1997.

[mawass-etal:eurospeech97] Khaled Mawass, Pierre Badin & Gérard Bailly
Synthesis of fricative consonants by audiovisual-to-articulatory inversion
European Conference on Speech Communication and Technology, vol 3, pages 1359--1362, Rhodes - Greece, 1997.

[morlec-etal:eurospeech97] Yann Morlec, Gérard Bailly & Véronique Aubergé
Synthesising attitudes with global rhythmic and intonation contours
European Conference on Speech Communication and Technolog , vol 1, pages 219--222, Rhodes - Greece, 1997.

[morlec-etal:jst97] Yann Morlec, Gérard Bailly & Véronique Aubergé
Apprentissage automatique d'un module de génération multistyle de l'intonation
1ères JST FRANCIL , pages 407--412, Avignon--France, 1997.

[morlec-etal:etrw97] Yann Morlec, Gérard Bailly & Véronique Aubergé
Generating the prosody of attitudes
ETRW Workshop on Prosody , pages 251--254, Athens - Greece, 1997.
[neagu-bailly:eurospeech97] Adrien Neagu and Gérard Bailly
Relative contributions of noise burst and vocalic transitions to the perceptual identification of stop consonants
European Conference on Speech Communication and Technology, vol 4, pages 2175--2178, Rhodes - Greece, 1997.
[rilliard-etal:jst97] Albert Rilliard, Véronique Aubergé, Gérard Bailly & Yann Morlec
Vers une mesure de l'information linguistique véhiculée par la prosodie
1ères JST FRANCIL, pages 481--487, Avignon - France, 1997.

[badin-etal:etrw98] PierreBadin, Gérard Bailly & Louis-Jean Boë
Towards the use of a Virtual Talking Head and of speech mapping tools for pronunciation training
ETRW on Speech Technology in Language Learning, Stockholm - Sweden, 1998.
[badin-etal:ETRW-Jenolan98] Pierre Badin, Gérard Bailly, Monica Raybaudi & Christophe Segebarth
A three-dimensional linear articulatory model based on MRI data
ESCA/COCOSDA International Workshop on Speech Synthesis, page 249--253, Jenolan Caves - Australia, 1998.
[allessandro-etal:ETRW-Jenolan98] Christophe d'Alessandro, Véronique Aubergé, Gérard Bailly, Frédéric Béchet, Philippe Boula de Mareuil, Jean-Philippe Goldman, Eric Keller, Vincent Pagel, Douglas O'Shaughnessy, François Yvon & Brigitte Zellner
Joint evaluation of text-to-speech synthesis in French within the AUPELF ARC-B3 project,
ESCA/COCOSDA International Workshop on Speech Synthesis, page 11-16, Jenolan Caves -Australia, 1998.
[badin-etal:mod3D-icslp98] Pierre Badin, Gérard Bailly, Monica Raybaudi & Christophe Segebarth
A three-dimensional linear articulatory model based on MRI data.
International Conference on Speech and Language Processing, page XXXXX, Sydney - Australia, 1998.
[badin-etal:mod3D-jep98] Pierre Badin, Laurent Pouchoy, Gérard Bailly, Monica Raybaudi, Christophe Segebarth, Jean-François Lebas, Mark Tiede, Eric Vatikiotis-Bateson & Y. Tohkura
Un modèle articulatoire tridimensionnel du conduit vocal basé sur des données IRM
Journées d'Etudes sur la Parole , pages 283--286, Martigny - Suisse, juin 1998.
[bailly:icp98] Gérard Bailly
Cortical dynamics and biomechanics
Bulletin de la Communication Parlée , 4:35--44, 1998.
[bailly:scom98] Gérard Bailly
Learning to speak. sensori-motor control of speech movements
Speech Communication , 22(2--3):251--267, 1998.
[bailly-etal:machoire-icslp98] Gérard Bailly, Pierre Badin & Anne Vilain
Synergy between jaw and lips/tongue movements : consequences in articulatory modelling
International Conference on Speech and Language Processing, page 417--420, Sydney - Australia, 1998.
[bailly-etal:machoire-jep98] Gérard Bailly, Pierre Badin & Anne Vilain.

Contribution de la mâchoire à la géométrie de la langue dans les modèles articulatoire statistiques.
Journées d'Etudes sur la Parole, pages 287--290, Martigny - Suisse, juin 1998.

[bailly-etal:cost258-98] Gérard Bailly, Eddy Bernard & Pierre Coisnon.
Sinusoidal modelling.
Cost258 Workshop, Vigo - Spain, 1998.

[morlec-etal:lrec98] Yann Morlec, Albert Rilliard, Gérard Bailly & Véronique Aubergé
Evaluating the adequacy of synthetic prosody in signaling syntactic boundaries: methodology and first results
First International Conference on Language Resources and Evaluation , Granada, Spain, 1998.

[neagu-bailly:icslp98] Adrien Neagu & Gérard Bailly
Collaboration and competition of burst and transition cues for the perception and identification of French stops.
International Conference on Speech and Language Processing, page 2127--2130, Sydney, Australia, november 1998.

[yvon-etal:csl98] François Yvon, Philippe Boula de Mareuil, Christophe d'Alessandro, Véronique Aubergé, Michel Bagein, Gérard Bailly, Frédéric Béchet, Salia Foukia, Jean-Philippe Goldman, Eric Keller, Douglas O'Shaughnessy, Vincent Pagel, Frédérique Sannier, Jean Véronis & Brigitte Zellner
Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French,
Computer Speech and Language, 12, pages 393--410, 1998.

[bailly:eurospeech99] Gérard Bailly
Accurate estimation of sinusoidal parameters in an harmonic+noise model for speech synthesis
European Conference on Speech Communication and Technology, vol 3, pages 1051--1054, 1999.

[holm-etal:icphs99] Bleike Holm, Gérard Bailly & Colette Laborde
Performance structures of mathematical formulae
International Congress of Phonetic Sciences , San Francisco, USA, vol 2, pages 1297--1300, 1999.

[morlec-etal:scom99] Yann Morlec, Gérard Bailly & Véronique Aubergé
Generating prosodic attitudes in French: data, model and evaluation
Speech Communication, (to appear).

[mawass-etal:aa] Khaled Mawass, Pierre Badin & Gérard Bailly
Articulatory synthesis of French fricative consonants
Acta Acoustica, (accepted).

[morlec-etal:eurospeech99] Yann Morlec, Gérard Bailly & Véronique Aubergé
Training an application-dependent prosodic model: corpus, model and evaluation
European Conference on Speech Communication and Technology , vol 4, pages 1643--1646, 1999.

ANNEXE 1 : ARCHITECTURE DES SYSTEMES DE SYNTHESE

Même si certains travaux ne se rapportent qu'indirectement à la thématique développée dans ce mémoire, il me semble important que certains travaux d'arrière-plan que j'ai encadrés soient mentionnés eu égard aux efforts fournis par les étudiants qui y furent impliqués et leur valeur de structuration intrinsèque de l'équipe.

Les rapports entretenus entre représentations phonétiques et synthèse sont nombreux et complexes : la synthèse a souvent été employée pour construire ou déformer des stimuli afin de sonder notre système perceptif ; la synthèse de parole en retour se nourrit des représentations phonétiques pour franchir le fossé existant les représentations symboliques construites par les traitements automatiques de la langue et les représentations continues des signaux de la communication parlée. Cette mise à disposition d'outils d'analyse et de validation ainsi que cette nécessité d'enrichir sans cesse notre connaissance des modèles que nous mettons en œuvre dans les systèmes de synthèse, cette synergie ne peut s'opérer que si les chercheurs sont acteurs dans les disciplines en amont et en aval de la connaissance. Ceci avait conduit Christian Benoît à énoncer le principe de base du fonctionnement de l'équipe synthèse : aucun membre permanent ne doit y émarger à 100% et l'équipe ne doit rien coûter au reste du laboratoire ! Il a tout le temps réussi à tenir ce pari…

Bâtir un système de synthèse (ou de reconnaissance) et lui assurer une pérennité en lui faisant subir les outrages des grands bouleversements informatiques et des changements des nombreux modèles de traitement de la langue écrite et orale impliqués dans de tels systèmes, est un défi quasi journalier. Cette tâche est souvent ingrate et de nombreuses initiatives visant à fournir des logiciels de base aidant à la construction de tels ensembles n'ont vu le jour que récemment [dutoit-etal:icslp96] [taylor-etal:etrw98]. Il faut d'ailleurs saluer cet effort même si dans le même temps il faut craindre l'uniformisation des formalismes que cette facilité peut engendrer !

Dans les années 80, il existait peu d'environnements permettant de construire des systèmes de synthèse tout en gardant une lisibilité relative des modèles, algorithmes et données utilisées. Seuls quelques projets de synthèse multilingues utilisaient des compilateurs de règles qui le plus souvent exploitaient soit une structure de données trop simple pour représenter de riches structures phonologiques [carlson-granstrom:icassp76] soit une structure de données trop flexible pour permettre une écriture simple des manipulations de structures élémentaires [hertz-etal:assp85].

De nombreux étudiants - souvent informaticiens de formation - ont donc ainsi construit par couches successives un outil de recherche et de valorisation des travaux décrits plus haut. Le nom, COMPOST, suggère plus un état de décomposition avancé qu'une technologie flamboyante. Le mot est cependant synonyme de vie et d'espoir…

Le cahier des charges

La construction d'un système de synthèse opérationnel nécessite la collaboration de nombreux spécialistes et l'intervention de plusieurs générations de chercheurs. Les disciplines impliquées vont du traitement automatique de la langue écrite (lexique, syntaxe, sémantique, phonétisation…) au traitement du signal et de l'image en passant par le traitement de la langue orale (phonologie, phonétique…). Toutes ces disciplines ont de plus recours à la psychologie expérimentale pour recueillir sur des sujets de référence les données indispensables aux modèles cognitifs constitutifs du clonage en gestation. De plus le système en cours de construction doit pouvoir lui aussi être un objet d'études afin de valider en permanence les options prises. Ces différentes disciplines représentent l'énoncé sous des

formes diverses et doivent contribuer à enrichir par de multiples connaissances a priori la maigreur de l'information intrinsèque d'une séquence de caractères alphanumériques. L'enjeu majeur auquel fait face le développement d'un système de synthèse à partir du texte est de permettre le développement, le test de modules spécifiques soit de manière autonome en maîtrisant la qualité des représentations d'entrée soit dans la chaîne de traitement complète de manière à pouvoir voir comment se propagent les erreurs générées par les modules en amont au travers le traitement. Un autre enjeu est de pouvoir offrir à la fois une représentation riche des données paralinguistiques, linguistiques et des signaux manipulés et de leur instanciation dans l'énoncé traité mais aussi contraindre suffisamment cette représentation de manière à assurer une lisibilité, une réusabilité et une simplicité d'utilisation maximale. Le dernier enjeu qui a présidé à l'élaboration de COMPOST est de disposer tout à la fois d'un système de développement de systèmes de synthèse multilingue satisfaisant les contraintes énoncées précédemment mais aussi d'un système de synthèse où les traitements proposés puissent être directement appliqués sur une entrée quelconque en temps réel.

Rapide aperçu

COMPOST a fait l'objet du travail de thèse de M. Alissali suite au travail initial d'A. Tran et de P. Marteau. De nombreux étudiants ont contribué ensuite à la version actuelle notamment I. Oueichek auquel on doit les divers gestionnaires et le quadruplet de DESS informatique qui a redessiné en langage objet la structure globale du serveur.

Architecture logicielle

COMPOST se reste un système original sans analogue dans la littérature : d'un point de vue architecture logicielle, il fonctionne sur le principe client-serveur. Le serveur se compose de quatre entités : un gestionnaire de sessions qui gère les requêtes, les options de synthèse retenues par chaque client, un gestionnaire de ressources qui alloue et partage les diverses ressources linguistiques utilisées par les divers modules de traitement (dictionnaires, paramètres d'un réseau de neurones ou d'une chaîne de Markov…), un gestionnaire de procédures algorithmiques multilingues pouvant être recrutées dynamiquement par toute chaîne de traitement (analyseur morphologique, modèle de n-grams, synthétiseurs…) et un moteur d'inférence chargé d'exécuter les divers traitements impliqués par chaque scénario.

Les scénarios

Un scénario décrit et gère l'ensemble des traitements impliqués par la synthèse d'un texte dans une langue donnée. Il comprend la définition des unités élémentaires de représentation de l'énoncé au cours de son traitement (caractères d'entrée : alphanumériques, signes…, phonèmes, syllabes, mots, syntagmes, propositions, phrases…), la définition de l'interface avec le client (ensemble de paramètres dynamiques conditionnant de manière générale un ou plusieurs traitements généralement non déductibles du texte : débit, caractéristiques paralinguistiques…) et un ensemble de blocks de traitements conditionnés par l'interface. Chaque bloc est constitué par une séquence d'appels à des procédures algorithmiques ou à des jeux de règles. Ces procédures et ces jeux de règles permettent de créer, modifier ou enrichir une structure arborescente dont les nœuds sont des instances des unités élémentaires définies plus haut.

Les unités élémentaires de représentation

Les unités élémentaires de représentation de l'énoncé sont des objets regroupés en classes naturelles (phonèmes, lettres, morphèmes, classes morpho-syntaxiques… mais aussi cibles pour décrire une trajectoire paramétrique) possédant des attributs communs (trait, entier, chaîne de caractères, réel…). Un même ensemble d'objets dits statiques, instanciés une seule

fois au chargement d'un scénario par un client, décrit l'interface. Les attributs de ces objets permettent alors de paramétrer les blocs, les appels procéduraux et les règles. Ainsi le même facteur débit peut affecter de multiples phases de traitement en conditionnant aussi bien l'application de règles d'élision du schwa que paramétrant un algorithme de distribution de durées segmentales dans une syllabe.

Les règles

Une règle élémentaire effectue une transduction d'arborescence sous contexte. Elle permet d'impliquer des objets ou des sous-arbres d'entrée ou des contextes gauche ou droit dans le sous-arbre de sortie. Elle permet en outre de conditionner l'application de la règle à des tests sur des expressions impliquant des attributs des objets du sous-arbre d'entrée, de modifier les attributs des objets du sous-arbre de sortie et d'impliquer dans ces opérations des attributs d'objets statiques. C'est par ce dernier mécanisme que les objets de l'interface conditionnent l'exécution des règles.

Bibliothèque algorithmique

A part la manipulation de connaissances linguistiques où les règles sont la solution la plus élégante et la plus efficace pour faire face à l'arbitraire du signe, la plupart des opérations de transduction de structures ou d'enrichissement - de décoration - de l'arbre de représentation du discours fait appel à des connaissances algorithmiques soit par le biais d'algorithmes de traitement des signaux soit par du traitement des formes. En effet, comme nous l'avons vu ci-avant les systèmes d'association sous-symboliques permettent d'élaborer des solutions optimales à des problèmes mal-posés et possédant des variables cachées en ayant souvent recours à une modélisation probabiliste et à un apprentissage hors-ligne de solutions type.

La fonction assurée peut être alors encapsulée dans une procédure à laquelle on spécifiera quelles informations (objets, attributs) de l'arbre d'entrée elle doit manipuler. Cette stratégie permet alors d'engranger un certain nombre de ressources réutilisables allant de l'analyse morphologique aux synthétiseurs courants de parole en passant par le filtrage syntaxique et la modélisation prosodique.

Quelques commentaires…

C'est grâce à cette architecture que COMPOST fût l'un des premiers (sinon le premier) synthétiseurs de parole à être présent sur le WEB. La structure et la réusabilité des composants algorithmiques de COMPOST a permis à deux étudiants (J. Camps et D. Sos) de développer en quelques mois des systèmes de synthèse complets du castillan et du catalan qui fonctionnent encore à présent. C'est enfin ce système qui a été utilisé pour implanter la synthèse par règles de M. Guerti et les divers systèmes de synthèse audiovisuelle développés par l'équipe animée par C. Benoît. La constitution de ce système a bénéficié d'un soutien européen (projet MultiWorks) ainsi que d'un contrat CNET.

Il reste que ce système a pour vocation non seulement de capitaliser notre savoir-faire et d'être une vitrine technologique mais aussi d'être un outil de recherches constamment remis en question pour s'adapter à de nouveaux objets d'études. La part de génie logiciel impliquée dans ces adaptations est importante et souvent lourde à gérer dans un laboratoire où l'informatique n'est pas une composante dominante.

ANNEXE 2 : DOSSIER D'HABILITATION

Curriculum Vitæ

- Né le 22 juillet 1959 à Besançon
- Adresse : ICP-INPG, 46 av. Félix Viallet, 38031 Grenoble CEDEX
- Tél : 04 76 57 47 11 - Fax : 04 76 57 47 10
- Email : bailly@icp.inpg.fr; URL: http://www.icp.inpg.fr/bailly.html

- Ingénieur ENSERG en 1981
- DEA d'Electronique INPG en 1981
- Doctorat d'Ingénieur INPG en 1983
- Coopérant scientifique à l'INRS Télécommunications - Montréal - Canada en 1984
- Attaché de recherches à l'INRS Télécommunications - Montréal - Canada en 1985
- Chargé de recherches au CNRS depuis juin 1986; CR1 depuis 1991

Administration de la recherche

- Membre élu du conseil de laboratoire de l'ICP depuis 1988
- Responsable de l'équipe Synthèse depuis 1995
- Membre suppléant puis titulaire de la CSE 61ème section en 1999
- Co-organisateur de la 1st International Conference on Speech Synthesis à Autrans en 1990
- Co-organisateur des Journées d'Etudes sur la Parole à Aussois en 2000
- Relecteur du Journal of Acoustical Society of America, Journal of Phonetics, Speech Communication, IEEE Transactions on Speech and Audio Processing et Acoustica

Publications

- Co-éditeur du livre *Talking Machines : Theories, Models and Designs*, chez Elsevier en 1992
- 9 articles (+2 soumis) dans des revues internationales
- 3 articles dans le Bulletion de l'ICP
- 10 chapitres de livres
- 58 articles dans des conférences internationales avec comité de lecture
- 6 articles dans des conférences sans comité de lecture

Enseignements

- Thèmes avancés à l'ENSERG de 1988 à 1996 « Communication parlée : du code au signal »
- Cours en Année Spéciale à l'ENSIMAG de 1988 à 1989 « Communication parlée »
- Cours de DEA à Lyon II en 1994 « Techniques et méthodes de synthèse de parole »
- Cours de Formation Continue à l'EPFL Lausanne en 1994 « Techniques et méthodes de synthèse de parole »
- Cours de DEA Sciences Cognitives de 1998 à 1999 « Synthèse de la parole »

Responsabilités de contrats et subventions

- Responsable du projet MULTIDIF de l'ACCT en 1989
- Responsable du contrat COMPOST avec le CNETen 1992
- Responsable de l'action Synthèse du projet Esprit II MultiWorks de 1989 à 1992

- Responsable de WP3 du projet Esprit III Speech Maps de 1992 à 1995
- Responsable de l'action « Codeurs » du thème B3 de l'AUPELF depuis 1995
- Responsable de l'action « Evaluation of Speech Synthesis Coders » du cost258 depuis 1996
- Responsable du contrat CNET « modèle 3D de visage parlant » de 1999 à 2000
- Co-responsable de l'action « Visiophonie & animation labiale scalable » du projet RNRT TEMPO-VALSE de 2000 à 2001

Encadrement

Thèses

- Coencadrement de la thèse INPG de O. Al Dakkak avec B. Guérin (1986-1988)

[aldakkak:these88] Omaima Al Dakkak

Extraction automatique de paramètres formantiques guidée par le contexte et élaboration de règles de synthèse

Thèse de troisième cycle, Spécialité Electronique, Institut National Polytechnique, Grenoble - France, 1988.

- Coencadrement de la thèse INPG de P.-F. Marteau avec M. Janot (1886-1988)

[marteau:these88] Pierre-François Marteau

Apport des notions de cibles et de trajectoires dans l'espcae acoustique en segmentation et étiquetage automatique du signal de parole

Thèse de troisième cycle, Spécialité Signal-Image-Parole, Institut National Polytechnique, Grenoble - France, 1988.

- Coencadrement de la thèse INPG de H.D. Wang avec D. Tuffelli (1988-1990)

[wang:these90] Haï-Dong Wang

Méthodologie de segmentation et d'étiquetage automatisés de bases de données acoustiques

Thèse de troisième cycle, Spécialité Signal-Image-Parole, Institut National Polytechnique, Grenoble - France, 1990.

- Directeur de thèse INPG de T. Barbe (1988-1990)

[barbe:these90] Thierry Barbe

Méthodologie et outils pour la mise en œuvre automatique d'une synthèse de parole de haute qualité

Thèse de troisième cycle, Spécialité Signal-Image-Parole, Institut National Polytechnique, Grenoble - France, 1990.

- Coencadrement de la thèse de P. Tripodi avec J.-C. Caërou (1990)

[tripodi:these91] Paul Tripodi

Etude et développement d'un serveur de messages à réponse vocale

Mémoire d'Ingénieur CNAM, Spécialité Electronique, Grenoble - France, 1991.

- Directeur de thèse de M. Guerti (1990-1993)

[guerti:these93] Manhia Guerti

Synthèse par règles du Français

Thèse d'Etat, Alger - Algérie, 1993.

- Coencadrement de la thèse de R. Laboissière avec J.-L. Schwartz et P. Perrier (1990-1992)

[laboissiere:these92] Rafaël Laboissière.

Préliminaires for une robotique de la parole: inversion et contrôle d'un modèle articulatoire du conduit vocal

Thèse de troisième cycle, Spécialité Sciences Cognitives, Institut National Polytechnique, Grenoble - France, 1994.

- Directeur de thèse de M. Alissali (1990-1993)

[alissali:these93] Mamoun Alissali

Architecture logicielle pour la synthèse multilingue de la parole

Thèse de troisième cycle, Spécialité Informatique, Institut National Polytechnique, Grenoble - France, 1993.

- Directeur de thèse de P. Barbosa (1992-1994)

[barbosa:these94] Plinio Barbosa

Caractérisation et génération automatique de la structuration rythmique du français

Thèse de troisième cycle, Spécialité Sciences Cognitives, Institut National Polytechnique, Grenoble - France, 1994.

- Directeur de thèse de Y. Morlec (1995-1998)

[morlec:these97] Yann Morlec

Génération multiparamétrique de la prosodie du Français par apprentissage automatique

Thèse de troisième cycle, Spécialité Sciences Cognitives, Institut National Polytechnique de Grenoble, Grenoble - France, 1997.

- Directeur de thèse d'A. Neagu (1995-1999)

[neagu:these98] Adrien Neagu

Analyse articulatoire du signal de parole: caractérisation des syllabes occlusive-voyelle en Français

Thèse de troisième cycle, Spécialité Signal, Image, Parole, Institut National Polytechnique, Grenoble - France, 1998.

- Directeur de thèse INPG de B. Holm (depuis 1998)

[holm:these] Bleike Holm

La prosodie de l'énonciation de formules mathématiques

Thèse de troisième cycle, Spécialité Sciences Cognitives, Institut National Polytechnique de Grenoble, Grenoble - France.

DEAs

DEA Sciences Cognitives :

- P. Barbosa (1991)
- V. Aubert en co-encadrement avec E. Castelli (1992)
- Y. Morlec (1994)
- A. Rilliard en co-encadrement avec V. Aubergé (1996)
- A. Capobianco (1998)
- B. Holm (1998)

DEA Signal-Image-Parole :

- J.P. Liu (1987 & 1988)
- J. Camps (1993)
- L. Roussarie en co-encadrement avec E. Castelli (1992)
- C. Vescovi en co-encadrement avec E. Castelli (1993)
- A. Neagu en co-encadrement avec J.-F. Sérignat (1994)
- D. Riquet (1997)
- E. Bernard (1998)

DESS Informatique

- L. Escat, L. Bollondi, X. Sirvent, D. Bagnol (1996)

Projets Ingénieur, Maîtrises

- A.Tran & P. Marteau, ingénieur ENSIMAG (1988)
- J.-P. Dezelus & S. Lunati, IUT Informatique Orléans (1989)
- F. Lanurien & xxxx, Maîtrise linguistique Paris 7 (1989)
- M. Olesen & M. Bach, ingénieur Université d'Aalborg - Danemark (1990)
- M. Vesterbacka, ingénieur Université de Linköping - Suède (1990)
- E. Lucazeau & L. Gavini, ingénieur ENSERG (1990)
- N. Robin, maîtrise U2-Grenoble (1990)
- J. Bonnyman, ingénieur Université de Nottingham - Angleterre (1991)
- J. Camps & D. Sos, ingénieur Université de Barcelone - Espagne (1992)
- I. Oueichek, ingénieur ENSIMAG (1992)
- L. Roussarie, ingénieur IEG (1992)
- C. Vescovi, ingénieur ENSERG (1993)
- P. Pianu & J.-F. Vire, ingénieur ENSERG (1994)
- C. Leclerq, ingénieur ENSERG (1996)
- T. Crépillat, maîtrise U3-Grenoble (1996) avec V. Aubergé
- R. Ducharme, maîtrise INRS Télécommunications - Canada (1996)
- D. Riquet, ingénieur ENSERG (1997)
- E. Bernard, ingénieur ENSERG (1998)
- P. Coisnon, ingénieur ENSIEG (1998)

Autres stages

IUT Informatique

- E. Petit, D. Po & D. Reboud , IUT2, Grenoble (1987)
- F. Balestra & B. Gilles, IUT2, Grenoble (1997)
- S. Aboufarid & V. Le Pottier, IUT2, Grenoble (1998)
- L. Arribat & Y. Calibet, IUT2, Grenoble (1999)

Stages ingénieur

- J. Camps, Université R. Lull, Barcelone-Espagne (1992)
- D. Sos-Vallès, Université R. Lull, Barcelone-Espagne (1992)
- J.L. Cherbonnier, Université du Maine, Le Mans (1993)
- M. Odisio, 2me année ENSERG (1999)

Séminaires et conférences invitées

- Séminaire Sciences Cognitives - Grenoble (1990) : « l'inversion en parole »
- Séminaire Sciences Cognitives - Grenoble (1990) : « Apprentissage de gestes articulatoires à partir du son » (avec R. Laboissière)
- Séminaire institut de Phonétique - Aix-en-Provence (1991) : « Génération automatique de la prosodie »
- Séminaire LIMSI- Paris (1992) : « Systèmes de synthèse de parole à partir du texte »
- Séminaire SFA-IRCAM - Paris (1993) : « Robotique et synthèse de parole »
- Séminaire DIST - Gênes (1993) : « Robotic approaches to speech generation »
- Jubilé Dominique Sabouraux - Rennes (1995) : « Contrôle moteur en parole »
- Ecole d'été GFCP - Lumigny (1995) : « Pistes de recherches en synthèse de parole »
- Séminaire Haskins - New Haven (1996) : «"Sensori-motor control of speech movements »

- ETRW Workshop on Speech Production - Autrans (1996) : "Sensori-motor control of speech movements »
- Séminaire AIMS - Stuttgart (1996) : « Control of speech movements »
- Journées ELESA « Machines communicantes » Grenoble (1998) « Les têtes parlantes »
- Cost258 Vigo (1998) « Sinusoidal modelling » & Lausanne (1999) « Signal generation test array »

Participations à des jurys de thèses

- H. Valbret, ENST (1995)
- F. Beaugendre, Orsay (1996)
- K. Mawass, INPG (1997)
- L. Reveret, INPG (1999)

ANNEXE 3 : PHOTOCOPIES D'ARTICLES

[barbosa-bailly:scom94] Plínio Barbosa & Gérard Bailly
Characterisation of rhythmic patterns for text-to-speech synthesis
Speech Communication, 15:127--137, 1994.

[bailly:atr97] Gérard Bailly
No future for comprehensive models of intonation?
Computing prosody: Computational models for processing spontaneous speech, Yoshinori Sagisaka, Nick Campbell & Norio Higuchi, editors, pages 157--164. Springer Verlag, 1997.

[bailly:scom98] Gérard Bailly
Learning to speak. sensori-motor control of speech movements
Speech Communication , 22(2--3):251--267, 1998.

[morlec-etal:sc99] Yann Morlec, Gérard Bailly & Véronique Aubergé
Generating prosodic attitudes in French: data, model and evaluation
Speech Communication, (to appear).