PERCEPTION DE LA PAROLE : INVARIANCE ET VARIABILITÉ

Jean-Luc Schwartz

COLLOQUE DE PHYSIQUE
Colloque C2, supplément au n°2, Tome 51, Février 1990 C2-461

1er Congrès Français d'Acoustique 1990

PERCEPTION DE LA PAROLE : INVARIANCE ET VARIABILITÉ

J.-L. SCHWARTZ

Institut de la Communication Parlée, INP/ENSERG, Université Stendhal UA CNRS n°368, 46 Avenue Félix Viallet, F-38031 Grenoble Cedex, France

Résumé - Partant d'une discussion générale sur les "deux problèmes de l'invariance" en perception, notamment en perception visuelle, nous discutons trois positions célèbres sur l'invariance en perception de parole : l'hypothèse de l'invariance acoustique, de articulatoire ou adaptative, et nous concluons sur l'intérêt d'une stratégie orientée vers les bas niveaux de traitement.

Abstract - Starting with a general discussion of the "two invariance problems" in perception, and mainly in visual perception, we discuss three famous statements about invariance in speech perception : the acoustic invariance assumption, the articulatory invariance assumption, and the adaptive invariance assumption, and we conclude about the interest of a low level strategy.

Introduction

La spécificité et la formidable efficacité de la parole comme media de communication vient d'abord de ce que le linguiste André Martinet a appelé sa "double articulation" : articulation du message en "unités de première articulation" ou unités de sens - ce que Martinet appelle les monèmes : atomes du lexique et de la morphologie - puis en "unités de deuxième articulation" ou unités distinctives, les phonèmes. On met ainsi en évidence une série de passages permettant de passer du message à transmettre à la chaîne "d'atomes du code linguistique", les phonèmes : niveaux sémantique (la spécification du message), lexical et syntaxique ("les mots pour le dire", et leur bon ordonnancement grammatical) et phonologique (la spécification du code). Il reste à passer au signal correspondant à la chaîne de phonèmes à transmettre. Ce passage du code au signal est géré par le niveau phonétique, et met en oeuvre les processus moteurs adéquats : gestion spatio-temporelle des "articulateurs" de la parole, c'est-à-dire la glotte et les articulateurs "supra-glottiques" - langue, lèvres, mâchoire, velum, larynx - responsables de la forme du tuyau dont on joue pour produire le son, le conduit vocal.

La théorie acoustique permet jusqu'à un certain point de prédire ou d'expliquer les caractéristiques du signal à partir des positions des articulateurs. A l'inverse, il devrait être possible de "remonter" du signal au code, et ce fut la base du "rêve" de la reconnaissance automatique de la parole dans les années 50-60. En réalité, le problème du "décodage acoustico-phonétique" - retrouver la chaîne des phonèmes à partir de l'analyse du signal - apparaît extrêmement complexe. Pour préciser la nature de cette complexité et présenter les grandes lignes du débat actuel dans ce domaine, je rappellerai d'abord le cadre général dans lequel le problème du décodage acoustico-phonétique doit être inscrit, et que l'on pourrait appeler le "problème de l'invariance" dans les systèmes perceptifs. Je discuterai, dans ce cadre général, quelques propositions de "stratégies vers l'invariance" issues principalement de travaux sur la perception visuelle. Je montrerai ensuite ce que le "problème de l'invariance" a de spécifique en perception de parole, avec des exemples tirés de l'analyse articulatori-acoustique de réalisations d'unités linguistiques. Enfin, sachant ce que le problème de l'invariance a de classique mais aussi de spécifique en perception de parole, je présenterai trois positions de référence sur la recherche de l'invariance pour le décodage acoustico-phonétique, en tentant de dégager en conclusion quelques éléments qui me semblent importants pour élaborer une stratégie de recherche dans ce domaine.

I. Problématique générale de l'invariance

I.1. Du physique au symbolique, la "condensation perceptive"

L'objectif du décodage acoustico-phonétique est d'attribuer une valeur linguistique à un signal acoustique. Ce passage du physique au symbolique - le "gap de Smolensky", si l'on veut se référer à ce saut qualitatif dont Smolensky fait l'analyse pertinente pour montrer comment les modèles de réseaux de neurones pourraient tenter de le combler - est au coeur de tout mécanisme perceptif.

La tâche essentielle d'un système perceptif est de donner au cerveau les moyens de s'approprier le monde physique, de dialoguer avec lui. Ce monde est mathématiquement un monde du continu, et un monde de l'infiniment variable. Ainsi, par exemple, l'image que "mon cerveau observe sur ma rétine" à l'instant précis où j'entre pour la première fois de mon existence dans la salle de conférence X de l'ICPI à l'occasion du premier Congrès Français d'Acoustique est une image pour moi radicalement nouvelle, totalement inédite, et elle n'est plus la même un instant plus tard, puisque ma position a changé, que certains objets ont bougé, que la luminosité a légèrement varié, etc...

Il me faut alors trouver le moyen de condenser l'information contenue dans cette image, afin de lui associer une forme - dans laquelle la variabilité aura été très fortement réduite - qui puisse me permettre de la mettre en relation avec ma connaissance présente du monde physique, et, partant de cette connaissance présente, d'entrer réellement en communication avec la scène extérieure dont ma rétine m'a transmis l'image.

I.2. Les deux "problèmes de l'invariance"

Restons dans le cadre de la perception visuelle. La "condensation perceptive" peut prendre plusieurs formes. Elle peut conduire à tout un ensemble de représentations discrètes (catégorisées) ou continues de la scène observée. *In fine*, je dois, à partir de mon image rétinienne, résoudre deux problèmes :

* retrouver les objets physiques, stables, à partir de l'impression perpétuellement variable qu'ils produisent sur mes capteurs ;
* nommer les objets, c'est-à-dire les identifier dans des catégories d'objets semblables.

Toutes différentes que soient ces deux tâches, elles correspondent l'une comme l'autre à la définition, à l'intérieur de l'ensemble - infini - de toutes les images rétiniennes possibles, de classes d'équivalence. Ainsi, si je veux identifier, dans la salle X où je viens d'entrer, le projecteur de transparents qui va me servir pour mon exposé, il me faut pouvoir définir la classe d'équivalence "images produites par le rétroprojecteur de la salle X" pour le premier problème, la classe d'équivalence "images produites par un rétroprojecteur quelconque" pour le deuxième problème. La "condensation perceptive" me permettant de faire communiquer le donné brut délivré par mes capteurs et la connaissance du monde physique dont je dispose passe par une schématisation de ce donné brut - un passage à une forme codée, symbolique - et s'effectue par la définition de deux types d'invariants, associés à chacune des relations d'équivalence définies ci-dessus.

I.3. Gibson, Marr, Poggio, et la "solution néo-réaliste"

Il se pose à ce point de la discussion le problème de la manière dont notre système peceptif détermine l'une ou l'autre de ces classes d'équivalence. La ligne de démarcation est, on le sait, entre "partisans du cognitif" et "partisans du perceptif", ou entre stratégies "top-down" (descendantes) et "bottom-up" (ascendantes). Elle est, en fait, entre ceux qui croient l'invariant imposé sur le signal par le cerveau, après calculs, raisonnements et inférences complexes, mise en jeu de processus prédictifs, de connaissances sémantiques, ..., et ceux qui croient à un invariant qui s'impose, qui vient du signal et "n'a plus qu'à" être détecté, repéré, identifié par les niveaux supérieurs. Elle est au fond entre les descendants de la tradition idéaliste et ceux de la tradition réaliste en philosophie, de Platon et Aristote jusqu'à Descartes, Leibnitz, Locke, Berkeley, et plus récemment entre behavioristes, expérimentalistes, Gestaltistes ou innéistes.

I.3.1. Les bases générales posées par Gibson

La position "néo-réaliste" la plus radicale a été exprimée parmi les psychologues contemporains par James Gibson (1966). Gibson pose très clairement l'existence dans toutes les projections rétiniennes d'un invariant associé aux objets physiques, la tâche du système visuel étant précisément d'explorer le monde physique pour faire progressivement émerger de l'ensemble des images correspondantes la structure des objets. L'invariant est donc défini par un ensemble de relations à l'intérieur des images, et cet invariant est défini à travers le jeu des transformations de l'image par modification du point de vue.

Citons quelques articulations de la pensée de Gibson dans "The senses considered as perceptual systems" (1966), pp.195-199.

"Whenever an observer moves, the array changes (...). As a stimulus for an eye, motion perspective is vastly more informative than static perspective, in that it specifies more about the geometrical layout of the environment. Let us make this assertion explicit. (With a stationary array) a distorted room could be substituted for the normal room without the observer being any the wiser (...). The perspective mapping could be identical for a family of rooms with different edge-and-corner layouts. (...) Distorted rooms with unpatterned walls and floors have actually been constructed in several psychology laboratories in recent years. (...) The observer of such an abnormal room gest an illusory perception of a normal room. (...) But the demonstration works only so long as the observer sees the distorted room with a motionless head and with only one eye. If he moves, or looks with two eyes, the actual layout is perceived.

et, finalement (pp.270-271) :

"(...) The perception of objects (...) always involves the detection of invariants under changing stimulation. The dimensions of transformation are separated off, and those that are obtained by action get distinguished from those that are imposed by events. The exploratory perceptual systems typically produce transformations so that the invariants can be isolated. And the action of the nervous system is conceived as a resonating to the stimulus information, not a storing of images or a connecting up or nerve cells".

Les trois points clé sont donc les suivants :

1. "L'invariant de premier type" (l'invariance de l'objet physique) existe dans le flot des images rétiniennes.
2. Il est produit par la multiplication des images des objets, obtenues par une stratégie perceptive active d'exploration.
3. Il est détecté directement, par résonance du système nerveux sur l'information spécifique du stimulus, donc sans intervention de modules cognitifs complexes.

Cette position est précisée par Johansson (1984) dans les termes suivants :

" (...) The visual system in an automatic and unavoidable way analyses the optical flow in a corresponding set of interrelated rigid motions. (...) The process is supposed to be of an automatic character without any cognitive elements. (...) This theory (...) establishes (...) that there generally exist one-to-one relations between proximal stimulus and the concommitant percept.

I.3.2. L'entrée dans le monde computationnel avec Poggio et Marr

Les remarquables travaux de David Marr et Tomaso Poggio sur les traitements visuels doivent être considérés en référence à cette philosophie générale de Gibson.

Marr et Poggio adhèrent totalement au point 3 et partiellement au point 1 : il existe bien, pour chacun d'eux, une considérable capacité de traitement à bas niveau, traitements purement ascendants - il faut, avant de faire appel aux hauts niveaux, "tirer tout ce que l'on peut de l'image rétinienne" - permettant, sinon de détecter une invariance constitutive de l'objet perçu, du moins d'accéder à une représentation proche de ses caractéristiques physiques.

Citons Poggio (1984) qui rend ce point de vue très clair.

" One of the best definitions of low-level vision is that it is inverse optics. Most of the goals of low-level vision can be seen as the solution to inverse problems. Consider, for instance, the problem of recovering the threedimensional structure of a scene from the images of it. While in classical optics the problem is to determine the images given certain physical objects, we are confronted here with the inverse problem of finding their three-dimensional shape (and perhaps their physical properties) from the light intensity distribution in the image."

Par contre, le point 2 est incontestablement le plus faible, rien n'étant dit par Gibson sur la forme prise par ces invariants par transformation ni sur la manière dont le système nerveux "résonne" sur leur structure. Poggio insiste alors sur le rôle crucial joué en perception visuelle par les contraintes, les hypothèses de régularité sur monde physique.

Poggio (1984) : *"In 1923 Hadamard described the class of problems with which mathematical physics had to deal. He was concerned with what we could call now dynamical systems (...). He defined the problem to be well-posed when the solution satisfied these three conditions : (a) the solution exists, (b) the solution is unique, (c) the solution depends continuously on the initial data. (...) Inverse problems are usually ill-posed. Usually inverse problems are obtained from the direct problems by exchanging the role of solution and data. A typical example of inverse problem is the determination of the shape of a drum from its frequency of vibration, a problem which was made famous by Marc Kac. (...) I argue that most problems in vision are ill-posed problems in the sense of Hadamard. (...) The basic idea of regularization methods for this type of ill-posed problems is to restrict the functional space of the acceptable solutions by imposing a set of constraints."*

L'introduction de contraintes pour pallier le défaut de dimensionalité dans les problèmes d'inversion en perception visuelle a conduit à plusieurs succès éclatants : récupération de la forme tridimensionnelle à partir du mouvement (Ullman, 1979), de la forme à partir de l'ombre (Horn, 1975 ; Terzopoulos, 1984), détection de contours (Torre & Poggio, 1984), vision stéréoscopique (Marr & Poggio, 1976 ; Marr, 1982), ...

Comme le note Donald Hoffman (1984), *"Le système est 'contraint' de choisir l'interprétation la plus crédible en fonction des règles et des régularités. La règle d'inférence du système visuel serait donc fondée sur une loi (la projection) et une régularité (la nature rigide des objets)."*

Un autre point essentiel soulevé le plus explicitement par David Marr (1982) est le problème dans l'approche de Gibson de l'absence de quelque proposition que ce soit sur la nature des traitements permettant d'extraire de l'image rétinienne une connaissance des objets physiques. Marr insiste sur la nécessité de travailler dans un cadre de représentation des objets, et propose toute une série de traitements et de représentations hiérarchiquement organisées permettant précisément le passage d'une image 2D à une connaissance 3D complète. L'apport crucial de Marr - relayé clairement par Poggio - est d'insister sur le fait qu'il faut, même dans le cadre d'une psychologie néo-réaliste comme celle de Gibson, une connaissance approfondie des capacités de traitement du système nerveux, c'est-à-dire de son anatomie et de sa physiologie. A ce titre, en introduisant la notion des trois niveaux auxquels se pose tout problème de traitement humain de l'information (le niveau du "pourquoi faire", de l'objectif ultime des traitements ; celui du "comment", des algorithmes à mettre en oeuvre ; celui du "avec quoi", du substrat physiologique), et en posant la nécessité de les aborder de manière séparée mais interdépendante, Marr est sans doute le chercheur contemporain qui aura posé avec le plus de clarté le rapport entre une psychologie néo-réaliste et la nécessité - initialement plutôt défendue par le "camp néo-idéaliste" - d'une connaissance approfondie du cerveau humain, de la physiologie et de la psychophysiologie.

II. La spécificité du "problème de l'invariance" en perception de parole

La difficulté du décodage acoustico-phonétique tient à ce que, naturellement, le signal "correspondant" à un phonème donné "varie". Les causes de cette "variabilité" sont multiples. Nous allons d'abord chercher à déterminer celles qui peuvent être considérées comme relevant classiquement de l'un ou l'autre des deux "problèmes de l'invariance" que nous avons mis en évidence en perception visuelle, puis nous verrons en quoi la nature de la variabilité diffère essentiellement en perception de parole de ce qu'elle est en perception visuelle.

II.1. Variabilité non spécifique

Premier problème : la nature de l'inversion

Peut-on poser l'existence d'un système de perception de parole "bas niveau" dans les termes où Gibson et plus encore Poggio posent ceux d'une "perception visuelle bas niveau"?

Une version première de cette analogie consiste à remarquer que, entre le son émis par un locuteur et la réception par l'oreille de l'auditeur, se situent les modifications dues au canal de transmission, et que le premier problème à résoudre pour le système auditif est de séparer, de déconvoluer, ce qui, dans le signal capté, est dû à l'émission - la cause de l'excitation auditive - et ce qui est dû à la transmission et qui informe l'auditeur sur la position de la source par rapport à son oreille, position au sens large : localisation spatiale ou nature du canal de transmission. Dans ce cadre, "l'objet distal" à retrouver derrière le stimulus proximal qu'est le signal capté par le tympan serait donc "l'objet sonore" produit par le locuteur. Le problème se pose alors jusqu'à un certain point dans les mêmes termes pour l'audition et la vision, et les spécialistes de la localisation auditive par exemple savent bien que pour combler le déficit de dimensionalité entre la réception et le couple (source, transmission), il leur faut récupérer de l'information par diverses voies (exploration perceptive par modification de l'orientation du pavillon, utilisation d'éléments secondaires comme la durée de réverbération pour récupérer la distance, ... Ce problème est cependant marginal en perception de parole, et j'introduirai ultérieurement (voir III.2) une seconde version possible du problème de l'inversion, beaucoup plus pertinente dans notre domaine.

Second problème : la catégorisation des objets sonores

On peut, comme dans le cas précédent, mettre en évidence des problèmes d'identification de timbre qui relèvent du même type de problème que celui de la catégorisation des objets visuels (peu abordé jusqu'à présent). Bien que la nature des processus d'identification phonétique soient nécessairement spécifiques par l'existence même d'un code, ce domaine est d'ailleurs un problème général de la perception auditive, comme le montrent avec force les travaux de Pierre Schaeffer (1966) sur la perception des objets musicaux (voir ses hypothèses sur la notion d'instrument de musique, débouchant par exemple sur ce qu'il nomme la "loi du piano", permettant d'identifier la nature d'un son donné comme étant ou non émis par un piano).

La part de variabilité du signal associé à un code donné relevant exclusivement de ce point est contenue essentiellement dans les différences entre locuteurs :

(i) différences physiques, interlocuteurs (forme et longueur du conduit vocal ou des cordes vocales) ou même intralocuteur (voix parlée, chuchotée, criée, ...), se traduisant par des différences systématiques sur les caractéristiques spectrales du signal ; il faut, pour

en tenir compte, essayer de se doter de mécanismes de "normalisation";

(ii) variantes linguistiques ("individuation" du système phonologique, sélection de certaines stratégies articulatoires ou acoustiques) se traduisant par une multiplicité de chemins du code au signal, et imposant donc de connaître tous ces chemins pour tenter de remonter du signal au code.

II.2. Variabilité spécifique : coarticulation

Néanmoins, toutes ces causes de variation ne remettent pas en cause la notion même de correspondance entre un élément de la chaîne phonétique et une portion du signal. Il n'en est pas de même pour les **mécanismes de variabilité liés au temps**, que l'on peut regrouper sous le terme général de **"coarticulation"**. A leur base, il y a un principe commun : renforcer l'efficacité de la communication parlée en se donnant les moyens de réduire les efforts d'articulation tant que l'on maintient dans le signal une quantité d'information suffisante pour l'auditeur à qui l'on s'adresse. On peut définir deux mécanismes principaux de coarticulation.

(i) La "réduction vocalique" traduit le fait que les mouvements des articulateurs vers une "cible articulatoire" qui permettrait de produire un spectre correspondant à une voyelle donnée sont le plus souvent "insuffisamment énergiques" pour que cette cible soit atteinte dans le temps imparti. La trajectoire vers la cible est interrompue pour aller vers la cible suivante, et cependant les caractéristiques de cette trajectoire sont suffisantes pour que l'auditeur identifie la cible sans ambiguïté.

(ii) D'autre part, les phonèmes sont "coproduits", ce qui signifie que les mouvements des articulateurs pour réaliser un élément de la chaîne phonétique sont effectués en fonction des phonèmes antérieurs et postérieurs. Ces stratégies d'anticipation et de persévération sont rendues possibles par le fait qu'il y a tout un ensemble de configurations articulatoires possibles pour réaliser une configuration acoustique donnée : on peut alors se déplacer dans cet espace de configurations (cette "fibre" du passage de l'articulatoire à l'acoustique) de manière à minimiser les déplacements articulatoires pour les cibles voisines.

Réduction et coproduction ont donc un effet commun : toute portion du signal acoustique dépend non pas d'un seul élément de la chaîne phonétique, mais de tout un ensemble d'éléments, et cette dépendance est elle-même paramétrée par des facteurs non linguistiques : le temps et les caractéristiques d'articulation (certaines situations de communication imposent des contraintes d'"hyper-articulation", d'autres permettent au contraire une "hypo-articulation", pour reprendre les termes de Lindblom que nous présenterons plus loin, voir III.3.3.).

II.3. Une invariance négociée

On voit donc que ce qui fait la spécificité de la variabilité donc du problème de l'invariance dans le décodage acoustico-phonétique est d'abord liée à la nature temporelle du signal acoustique véhiculant l'information, et l'on insiste en général sur cette différence entre un signal acoustique que l'on ne peut analyser "hors du temps" et un signal optique permettant une analyse statique ou quasi-statique.

En fait, il y a, au-delà de cette différence de nature entre deux signaux physiques porteurs d'information vers le système perceptif, une autre différence de nature, mais cette fois entre les objets à percevoir dans l'un ou l'autre cas : la perception visuelle traite d'objets physiques, là où la perception de parole doit s'appliquer à des signaux biologiques émis par un système aussi complexe que le système de traitement qui les analyse, puiqu'il s'agit dans les deux cas de parties du cerveau humain.

L'existence de la coarticulation montre que le décodage acoustico-phonétique se situe au coeur d'un ensemble d'exigences de chacun des deux systèmes, exigences d'économie pour le système de production, exigences de compréhension pour le système de perception, et on peut s'attendre à ce que "l'invariance phonétique" dans le signal acoustique soit, si elle existe, une invariance négociée. Nous reviendrons en détail sur cette idée au paragraphe III.3.

III. Les stratégies vers l'invariance en perception de parole

Un panorama général des propositions théoriques et des résultats expérimentaux est bien sûr tout à fait hors de portée dans le cadre de cet article. Pour se faire cependant une idée des grands enjeux, des différents angles d'attaque possibles et des principales conceptions qui s'affrontent, j'ai choisi, assez arbitrairement, de présenter trois positions défendues avec une grande constance et une très solide argumentation théorique et/ou expérimentale depuis plus de 20 ans par des chercheurs qui comptent parmi les plus importants du domaine de la perception de parole - et plus encore de la parole en général. Ces positions derrière lesquelles se retrouvent dans chaque cas un large ensemble d'équipes et de travaux ont le double avantage d'être claires et tranchées - et, par corrolaire, risquées - et de constituer de bons phares pour s'orienter dans le paysage qui est le nôtre.

III.1. La "solution Stevens" : l'invariance acoustico-perceptive

Ken Stevens a fait, depuis le début des années 70, une série de propositions sur des propriétés du signal capables selon lui de signaler directement une catégorie phonétique, c'est-à-dire d'éliminer l'ensemble des causes de variabilité, qu'elles soient liées à la variabilité du locuteur, du style d'élocution ou du contexte phonétique. Je présenterai la philosophie générale de Stevens et des travaux reliés à cette philosophie autour de 4 points :

1. le "révélateur" fourni par les hypothèses de Stevens et Sheila Blumstein sur la détection du lieu d'articulation des plosives ;
2. le cadre général sous-jacent contenu dans la fameuse "Théorie quantique" ;
3. les hypothèses sur l'organisation du système auditif dans ce cadre ;
4. la position qui se déduit des points 1 et 2 sur la nature-même des systèmes phonologiques.

III.1.1. Un invariant acoustique pour le lieu d'articulation des plosives

L'étude de la perception du lieu d'articulation des plosives de l'anglais-américain conduit d'abord Blumstein, Stevens et Nigro (1977) à proposer l'existence de détecteurs associés à des propriétés élémentaires de diverses caractéristiques du signal acoustique correspondant, telles que: spectre de l'explosion, direction et vitesse des transitions de formant suivant l'explosion. Une série d'expériences sur l'adaptation - la manière dont l'exposition répétée à un signal modifie la catégorisation d'autres signaux plus ou moins voisins du signal adaptateur - induite par des signaux combinant des propriétés cohérentes ou contradictoires entre elles conduisent les auteurs

à confirmer l'existence de tels détecteurs de propriétés, dont les réponses devraient ensuite être intégrées pour fournir un critère global d'identification de la plosive.

Peu de temps après, Stevens et Blumstein (1978) ont franchi un pas important dans leur hypothèse sur l'invariance. Ils tentent en effet de montrer que la perception du lieu d'articulation des plosives /b,d,g/ à l'intérieur de stimuli de synthèse de type CV (Consonne-Voyelle), où la voyelle est l'une quelconque des voyelles /a,i,u/, est directement contrôlée par un seul paramètre perceptif qui est la forme d'ensemble du spectre calculé sur une portion du signal allant de l'explosion aux premières périodes du voisement vocalique. On aurait ainsi, à partir de ce calcul de spectre :

* un invariant "plosive bilabiale" (/b/) pour un spectre "diffus" (pas de pic net dans la zone 1-3 kHz) et "descendant" ou "grave" (avec A(F4)≤A(F3)≤A(F2), F2, F3, F4 étant les 2ème, 3ème et 4ème pics spectraux et A(Fi) leurs amplitudes) ;
* un invariant "plosive dentale" (/d/) pour un spectre diffus "montant" ou "aigu" (A(F4)≥A(F3)≥A(F2)) ;
* un invariant "plosive vélaire" (/g/) pour un spectre "compact" (un pic proéminent dans la zone 1-3 kHz).

Blumstein et Stevens (1979) confirment l'existence de ces invariants du signal acoustique par l'analyse de parole naturelle, en montrant qu'un algorithme d'identification du lieu d'articulation de plosives fonctionnant sur ce principe fournissait de bons scores de reconnaissance, indépendamment du locuteur (4 hommes, 2 femmes), du mode d'articulation (plosives voisées ou non), de la position dans la syllabe (initiale - CV- ou finale - VC) et du contexte (/i,e,a,o,u/ en position finale dans des syllabes CV, ou /i,ɛ,a,ʌ,u/ en position initiale dans des syllabes VC). Stevens et Blumstein en arrivent donc à l'hypothèse de l'existence d'invariants dans le signal acoustique, invariants capables de faire passer directement du signal au code, et ce par un calcul simple susceptible d'être effectué par des procédures purement ascendantes - et donc, en un certain sens, non cognitives. Leur position ultérieure sur la nature précise de l'invariant associé à chacun des lieux d'articulation se modifiera par la suite. Ainsi, Blumstein (1983) passe d'un invariant statique à un invariant dynamique, reposant sur une comparaison du spectre de l'explosion et du spectre des premières périodes de voisement - rejoignant ainsi la position de Kewley-Port (1980, 1983), insistant constemment sur la nécessité de définir l'invariance à partir d'analyses dynamiques. Néanmoins, l'hypothèse de fond ne change pas.

Ainsi, dit Blumstein (1983) : *"(...)There is acoustic invariance in the speech signal corresponding to the phonetic features of natural language. That is, it is hypothesized that the speech signal is highly structured in that it contains invariant acoustic patterns for the phonetic dimensions of language relating in particular to linguistic segments and to phonetic features, and that these poatterns remain invariant across speakers, phonetic contexts, and languages".*

III.1.2. La théorie quantique

Si de tels invariants "phonologiques" peuvent exister directement dans le signal acoustique, c'est d'abord, propose Stevens dès 1972 avec la première présentation de sa théorie quantique, grâce aux non-linéarités du passage de l'articulatoire à l'acoustique. L'idée de base en effet est qu'il existe dans un certain nombre de cas de production de parrole des conditions dans lesquelles une modification d'un paramètre articulatoire a sur un paramètre descriptif du signal acoustique une action fortement non linéaire telle que celle schématisée sur la Figure 1. Dans les domaines de variation I et III du paramètre d'entrée (articulatoire), les variations du paramètre de sortie (acoustique) sont faibles - zones dites de stabilité. Au contraire, dans le domaine II - zone d'instabilité - les variations du paramètre acoustique sont très rapides. Plus encore, Stevens (1989) insiste dans sa plus récente exposition de la théorie quantique sur le fait qu'il y a en général, dans ce schéma, une véritable transformation qualitative du paramètre acoustique de la zone I à la zone III, donc à travers la zone II.

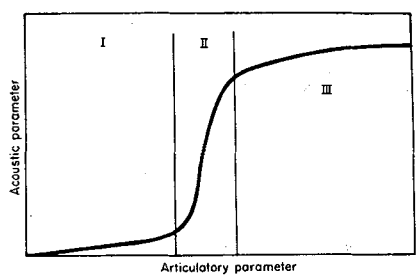

Figure 1 - Non-linéarité du passage d'un paramètre de commande articulatoire à un paramètre de sortie acoustique dans le paradigme de base de la Théorie Quantique (d'après Stevens, 1989)

Dans l'article central d'un numéro du *Journal of Phonetics* (N°17, Vol.1-2, 1989) consacré à sa théorie quantique, Stevens propose tout un ensemble de non-linéarités du passage de l'articulatoire à l'acoustique présentant de telles caractéristiques. Bien que les trois invariants qu'il proposait avec Blumstein à la fin des années 70 pour le lieu d'articulation des plosives n'y figurent pas précisément sous cette forme, le principe de recherche d'invariants acoustiques est clairement relié à cette dialectique stabilité/instabilité qui est le paradigme central de la théorie quantique, et qui explique le choix par Stevens de son nom, puisqu'il s'agit d'une théorie de la production de discontinuités dans un continuum physique.

III.1.3. Le rôle du système auditif dans le cadre proposé par Stevens

* Delgutte et le "renforcement de l'invariance" dans le système auditif périphérique

Dès la présentation initiale de sa théorie en 1972, Stevens fait l'hypothèse qu'il existe également des non-linéarités du type plateau/instabilité/plateau dans le passage de l'acoustique à l'auditif, c'est-à-dire que le système auditif est capable de renforcer la tendance présente au niveau acoustique.

Les exemples de non-linéarités du passage de l'acoustique à l'auditif sont peu convaincants dans l'article de 1972. Ils le sont da-

vantage dans l'article de 1989 bien que le noyau dur des arguments de Stevens soit incontestablement situé dans le passage de l'articulatoire à l'acoustique. Néanmoins, bien que les travaux de Bertrand Delgutte (1980, 1982, 1983) sur le codage de la parole dans le système auditif périphérique ne soient pas directement inscriptibles dans le cadre de la théorie quantique, ils fournissent très certainement les exemples les plus convaincants de cas où un mécanisme physiologique donné - l'adaptation nerveuse - bien décrit expérimentalement et théoriquement par les physiologistes permet un "renforcement" de l'invariance acoustique et à ce titre ils doivent être cités ici. Il faut d'ailleurs dire à quel point la synthèse entre les connaissances du grand expérimentateur et spécialiste du nerf auditif qu'est Nelson Kiang et celles du grand théoricien de l'invariance et de la parole en général qu'est Ken Stevens a été réussie par Delgutte.

Citons trois exemples tirés des travaux de Delgutte, montrant à chaque fois le rôle de l'adaptation nerveuse dans le codage de la parole dans le nerf auditif.

1. Une "meilleure invariance" pour le lieu d'articulation des plosives

Cette proposition ne vient en réalité pas de Delgutte lui-même mais d'une lecture que font Stevens et Blumstein de ses travaux sur l'adaptation. Considérant en effet que "l'invariant plosive dentale" avec la forme diffus descendant du spectre à l'explosion ne tient pas aussi bien pour /n/ (avec un spectre plutôt plat ou montant) que pour /d,t/, ils proposent (voir Blumstein et Stevens, 1979 ; Stevens, 1980) que le murmure nasal basse fréquence précédant, dans toute consonne nasale, l'explosion due à la détente dans le conduit oral, pourrait masquer partiellement la partie basse fréquence du "spectre auditif" à la détente, et rendre ainsi le poids relatif des hautes fréquences plus important, et la forme générale plus en accord avec leur hypothèse.

2. Une intégration auditive de propriétés acoustiques

Delgutte (1982) montre sur un modèle du système auditif périphérique comment, sous l'effet de l'adaptation nerveuse, deux types de modification du signal acoustique qui permettent sur des stimuli synthétiques de passer d'une fricative /ʃ/ à une affriquée /tʃ/ - une augmentation du temps de silence entre la voyelle précédente et le bruit de friction, ou une diminution du temps de montée de ce bruit de friction - ont une même action sur la réponse des fibres du nerf auditif, c'est-à-dire un renforcement du pic de décharge au début du bruit de friction, donnant naissance à la perception de l'explosion de l'affriquée.

3. Un renforcement du contraste entre plosives voisées et non voisées

Delgutte (1982) montre également comment une différence quantitative de VOT - Voice Onset Time, durée séparant l'explosion de la consonne du début du voisement de la voyelle qui suit, durée plus longue pour les plosives non voisées que pour les plosives voisées - peut conduire à une différence qualitative de réponse des fibres : il obtient en effet pour certaines fibres de son modèle deux pics de décharge pour une plosive non voisée et un seul pic pour une plosive voisée, ces fibres n'ayant pas le temps de récupérer de leur forte décharge à l'explosion pour manifester un nouveau pic de décharge au début du voisement dans le cas de plosives voisées.

Nous avons, dans notre laboratoire, poursuivi les investigations dans la voie ouverte par Delgutte, en montrant (Wu et al., 1988, 1989) comment un modèle du système auditif périphérique prolongé par un modèle de cellules "on" et "off" (cellules ne répondant qu'à un début ou une fin de stimulus) était capable de détecter avec efficacité des discontinuités spectro-temporelles du signal de parole (des "événements") reliés à des mécanismes de base de la production de parole (début et fin de vibrations laryngées, début et fin du bruit d'explosion ou de friction, début et fin d'état vocalique du conduit vocal).

*** Les détecteurs inspirés de la neurophysiologie de la vision**

Plus généralement, tout le travail de Delgutte vise à mettre en évidence des caractéristiques de réponse du nerf auditif différentes selon la catégorie phonétique du stimulus d'excitation. Ceci prépare la voie à ce qui, pour Stevens, doit être le niveau suivant dans le traitement auditif du signal acoustique : l'existence de véritables détecteurs rendus célèbres en neurophysiologie depuis les travaux de Hubel et Wiesel sur la physiologie de la vision, détecteurs de propriétés puis détecteurs d'invariance (fonctionnant directement sur le signal ou à partir de propriétés élémentaires) qui fourniront ainsi l'indication de la catégorie phonétique associée.

III.1.4. Conclusion : contraintes perceptives sur les systèmes phonologiques

L'idée générale de la théorie quantique est claire : les non-linéarités de la transformation articulatoire -> acoustique ajoutées au non-linéarités inhérentes aux traitements auditifs permettent de passer d'un continuum articulatoire à des catégories auditives qui fourniront la base des invariants phonologiques. L'essentiel de la thèse de Stevens est, à partir de là, que les systèmes phonologiques se sont précisément constitués autour de ces catégories auditives, qui fournissent un support physique (physiologique) à la théorie des traits distinctifs proposée d'abord par Jakobson, Fant et Halle (1963), puis par Chomsky et Halle (1968). Ainsi, deux points essentiels doivent être soulignés pour définir la position de Stevens dans le domaine de la parole :

* Stevens est un "réaliste", dans la mesure où il croit très fermement à une organisation de la "pâte acoustique" en catégories naturelles qui sont à la base de l'existence et de la perception du code phonologique. Il est important de citer à ce propos l'important ouvrage consacré par Jean Petitot (1985) à la position théorique de Stevens en relation avec les problèmes épistémologiques centraux que sont la réalité physique du discontinu et des catégories et leur rapport avec les catégories mentales, ainsi qu'avec le cadre général de la théorie des catastrophes de René Thom. Il faut noter également que les concepts déterminants que sont chez Stevens la recherche de discontinuités physiques et de traitements physiologiques adaptés à ces discontinuités (c'est-à-dire capables de les détecter) en font incontestablement un des représentants du domaine de la communication parlée les plus proches des positions défendues par David Marr dans le domaine de la perception visuelle.

* Cependant, il est clair qu'une seconde specificité de Stevens est qu'il croit à un invariant dans le signal acoustique lui-même et non dans le monde physique des causes de ce signal, le monde de l'articulatoire, au contraire de la position que nous allons décrire dans le paragraphe suivant. Le point crucial est que, pour Stevens, toute l'organisation du système phonologique et des mécanismes de production est orientée vers la sortie, c'est-à-dire vers l'auditeur, et que les contraintes majeures sont ainsi des contraintes du système de perception, le système auditif.

Notons enfin que nous avons eu nous-mêmes l'occasion de mettre en évidence la fertilité des hypothèses de Stevens, en :

- produisant des données expérimentales assez fortes en faveur de l'existence dans le système de traitement perceptif des spectres

vocaliques d'un mécanisme d'intégration à large bande (Schwartz et Escudier, 1987, 1989) ;
- montrant comment ce mécanisme était capable, dans le droit fil de la théorie quantique, de "produire des discontinuités" sur un continuum acoustique (Abry et al., 1989) ;
- et, enfin, le rôle qu'il était susceptible de jouer dans l'organisation des systèmes vocaliques (Schwartz, 1987 ; Abry et Schwartz, 1987 ; Schwartz et al., 1989).

III.2. La "solution Liberman" : l'invariance dans le geste articulatoire

III.2.1. Les principes de la théorie motrice

Tout comme Stevens, Liberman et les principaux chercheurs qui ont partagé et développé ses thèses - Cooper, Fowler, Mattingly, Shankweiler, Strange, Studdert-Kennedy, Verbrugge..., la plupart provenant des célèbres "Laboratoires Haskins" - croient à une invariance spécifiée quelque part dans le signal acoustique, mais leur second point de départ est que, à l'opposé de Stevens, ils ne pensent pas possible de trouver cet invariant dans le signal lui-même. Bien au contraire, l'ensemble des recherches menées dans les années 60 dans ce laboratoire les ont convaincu de la profonde non-invariance phonétique du signal acoustique ! Les principales manifestations de cette non-invariance sont revues par Liberman et Mattingly (1985) : (i) un large ensemble d'indices acoustiques sont associés à un même percept, (ii) un même fait acoustique peut être perçu différemment selon le contexte, (iii) deux faits acoustiques différents peuvent être perçus identiques à cause du contexte. L'interprétation de l'ensemble de ces faits par Liberman et coll. est que la réponse perceptive est donnée en fonction de la nature du geste articulatoire qui est à l'origine des caractéristiques du signal acoustique. Ainsi, (i) l'ensemble des indices acoustiques donnant naissance à un même percept seraient eux-mêmes produits par un même geste articulatoire, (ii-iii) un fait acoustique donné serait perçu en relation avec d'autres faits acoustiques en fonction du geste articulatoire qui les a produits.

On en arrive ainsi à une position typiquement Gibsonnienne selon laquelle le signal acoustique n'est qu'un medium façonné par une cause physique plus profonde, qui réside dans le système de production des sons par le conduit vocal : les gestes articulatoires (Liberman et al., 1967) ou, plus en amont encore, les commandes motrices ou structures coordinatives qui ont permis de programmer ces gestes (Liberman et Mattingly, 1985). On voit alors que le "premier problème de l'invariance" - le problème classique de la psychologie Gibsonnienne - n'est plus, comme nous le proposions au paragraphe II.1., de retrouver le signal acoustique émis par la source - le conduit vocal du locuteur - mais la nature-même de la source. En termes techniques, l'objet distal de la perception n'est plus, pour Liberman, acoustique - l'onde émise - mais articulatoire - le système de production de cette onde.

Si les points de départ de cette théorie sont séduisants à plusieurs titres, et notamment parce qu'ils permettent de tenir compte du fait que chaque auditeur est aussi un locuteur, et peut donc s'appuyer sur une connaissance parfaite des mécanismes de passage de l'articulatoire à l'acoustique, il reste (!) à montrer comment elle permet de progresser sur le difficile chemin vers l'invariance. Nous allons le discuter sur deux exemples très classique tirés de problèmes posés par la coarticulation.

III.2.2. Invariance articulatoire et perception dynamique des voyelles

Le premier fait de coarticulation que nous avons mentionné (II.2) est la réduction vocalique : la cible acoustique caractéristique d'une voyelle isolée n'est le plus souvent pas atteinte en parole continue. Le problème se pose alors de savoir comment peut être identifiée cette "cible non atteinte". Notons tout de suite que la distance entre la position acoustique théorique et la position réelle peut être tout à fait importante et apte à provoquer des confusions très nettes sur la nature phonétique d'une voyelle si on ne prend en considération que cette position atteinte. Ainsi, nous avons montré dans notre laboratoire qu'un /a/ ou un /ɛ/ en contexte /iVi/ fournissaient, pour différentes conditions de débit ou d'accentuation, des positions formantiques totalement imbriquées (Beautemps, 1989). Pire encore, Demany et Semal (communication personnelle) rapportent qu'un très léger déplacement de formants à partir d'une voyelle donnée V1 en direction d'une autre voyelle V2 suffit à évoquer la perception d'une diphtongue V1V2 bien que l'excursion formantique reste très largement à l'intérieur de la zone associée à la perception de V1 en statique.

C'est donc sur la trajectoire toute entière qu'il faut rechercher l'information. En 1963, Lindblom a formulé et testé un modèle dynamique simple de réduction vocalique dans lequel il faisait l'hypothèse que la position atteinte s'approchait de la cible selon une loi exponentielle en fonction de la durée de la voyelle (cible mieux atteinte pour des voyelles longues que pour des voyelles courtes), puis il a introduit en 1968 un facteur complémentaire relié à la force d'articulation (cible mieux atteinte pour des voyelles accentuées, et pour de la parole mieux articulée). Un tel modèle acoustique peut être précisément relié à un modèle de commande articulatoire tel que celui proposé dans notre laboratoire par Perrier et al. (1989) et faisant appel à des processus dynamiques du second ordre (masses / ressorts). On peut alors faire l'hypothèse - "à la Liberman" - que la dynamique d'évolution du modèle vers sa cible spécifie la position de cette cible même si elle n'est pas atteinte, puisque, par exemple, la donnée de 5 points consécutifs d'un système linéaire du second ordre à un seul degré de liberté suffit mathématiquement à spécifier les trois paramètres inconnus (raideur, frottement, cible) qui le caractérisent. Dans cet ordre d'idée, une étude préliminaire nous a permis de montrer que, pour le même corpus /a/ vs /ɛ/ en contexte /iVi/ décrit ci-dessus, des informations sur la vitesse de variation des formants entre la voyelle initiale /i/ et la cible /a/ ou /ɛ/ suffisaient à inférer la nature de cette cible (Beautemps, 1989).

Indépendamment du problème de la réduction vocalique, un ensemble de travaux - partis encore, pour l'essentiel, des laboratoires Haskins - ont visé à montrer que des voyelles en contexte consonantique pouvaient dans certains cas être identifiées plus aisément que des voyelles isolées, et cela d'autant plus qu'il s'agissait de voyelles produites par différents locuteurs non connus de l'auditeur (Strange et al., 1976 ; Fowler et Shankweiler, 1978 ; Verbrugge et al., 1979). Ces résultats, abondamment discutés, voire critiqués (Pisoni et al., 1979; Gottfried, 1979; Macchi, 1980 ; Diehl et al., 1981 ; Assman et al., 1982), sont interprétés comment étant le signe qu'il existe une information dynamique dans la consonne qui spécifie la cible intentionnelle de la voyelle qui suit. Plus profondément, il faut voir dans ces travaux une réminiscence, parfois explicite (Fowler et al., 1979), de la position de Gibson sur la nécessité d'une exploration dynamique de l'objet distal de la perception afin de récupérer l'invariance dans le stimulus proximal, l'objet distal étant ici, bien sûr, le geste articulatoire vers la cible de la voyelle. La difficulté de cette approche est cependant que l'objet distal est lui-même un objet dynamique.

III.2.3. Fowler et la perception directe

Le second fait de coarticulation dont nous avons parlé est la coproduction. La conséquence en est que, si les segments abstraits d'une séquence définie au niveau phonologique (les phonèmes) sont par nature discrets, statiques et indépendants du contexte, les segments produits apparaissent dans le signal acoustique en se recouvrant, partiellement ou totalement, et sous une forme dynamique et dépendante du contexte (Fowler, 1983). Pour comprendre dans ce cadre comment peuvent fonctionner des mécanismes de segmentation (récupération des segments phonétiques à partir du signal acoustique), Carol Fowler invoque les expériences de Johansson sur la perception visuelle, et notamment sur ce qu'il appelle la "perception vectorielle" (perceptual vector analysis). Johansson (1950, 1964) montre que la perception de configurations dynamiques est spontanément analysée en composantes cohérentes du point de vue de leur mouvement (des éléments affectés de vecteurs vitesse identiques sont regroupés en unités cohérentes) et que ceci peut conduire à une décomposition vectorielle d'un vecteur vitesse en composantes de base reliées chacune à la vitesse d'une des configurations perçues.

Ceci montre, selon Fowler, que le système de perception de parole peut de la même manière décomposer le signal acoustique, non pas en unités acoustiques (en "unités d'apparence") mais en unités articulatoires (en "unités de cause"), en prenant dans diverses parties du signal ce qui relève d'un geste articulatoire (geste vocalique, geste consonantique), et cela seul.

Le problème général posé alors par la position de Liberman et des partisans de la théorie motrice, ou de Fowler et des tenants de la théorie de l'action et de la perception directe, est de montrer comment peut se traiter le problème de l'inversion, c'est-à-dire de la récupération des gestes articulatoires (voire même, pour Liberman et Mattingly, des commandes articulatoires ou structures coordinatives) à partir du son. La résolution de ce problème passe très probablement par une avancée significative des recherches sur la production de parole et les modèles articulatoires. Mentionnons finalement à quel point cette position commune de Liberman et Fowler les rend proches de la position de Poggio sur la perception visuelle. Il suffit, pour s'en convaincre, de rappeler le problème d'inversion posé par le mathématicien Marc Kac, cité en exemple par Poggio, et décrit récemment dans La Recherche sous le titre : "Peut-on entendre la forme d'un tambour ?" (Fabre, 1988).

III.3. La "solution Lindblom" : l'invariance adaptative

III.3.1. Marr et Poggio, Stevens et Liberman : questions sur une divergence

Nous avons conclu les deux paragraphes précédents par une double proximité : celle de la "solution Stevens" avec la position de Marr, et celle de la "solution Liberman/fowler" avec la position de Poggio. Ce fait est assez frappant, si l'on met en correspondance la grande connivence de points de vue entre Marr et Poggio d'une part, la divergence des appréciations de Stevens et de Liberman ou Fowler sur la variabilité du signal acoustique d'autre part. Où est l'erreur ?

* Il est certain que l'on retrouve dans la position de Fowler et bien plus encore dans celle de Liberman et Mattingly avec la théorie motrice le défaut majeur des travaux de Gibson (mais pas de ceux de Poggio) : celui d'une théorie séduisante et s'intégrant bien à un ensemble d'arguments de diverses natures, mais qui ne s'est pas jusqu'à présent avérée capable de produire de réelles perspectives de recherche, qu'il s'agisse de propositions de traitements ou d'expériences.

* Cette critique ne peut être adressée à la théorie quantique et à la position générale de Stevens sur l'invariance acoustique. Il est vrai, par contre, que la manière dont sont traités par Stevens et ses proches les problèmes liés à la coarticulation n'est pas vraiment convaincante. Ainsi, l'existence d'un invariant du lieu d'articulation des plosives, tous contextes et toutes conditions d'articulation et de débit confondus, semble difficile à accepter ; de même, rien n'est dit sur le traitement des cibles non atteintes pour la perception des voyelles. Dans ces deux problèmes de dynamique clairement liés à des exigences du système de production, la stratégie de Liberman et Fowler semble au contraire plus attrayante.

* Un dernier point doit être mentionné. Les conditions du passage de l'articulatoire à l'acoustique sont exploitées par Liberman et Fowler pour traiter de ce que nous avons nommé en I.2 le premier problème de l'invariance, tandis que ce passage est uniquement considéré par Stevens sous l'angle de ses non-linéarités, et ce afin de traiter précisément du second problème de l'invariance, c'est-à-dire, finalement, du problème de l'identité perceptive des catégories phonologiques, traits ou phonèmes. A ce titre, incontestablement, quelque chose de plus est dit par ce dernier sur la phonologie, et on sent a contrario une faiblesse dans ce domaine du côté de la théorie motrice ou de la théorie de l'action, comme le reconnait d'ailleurs Carol Fowler (1986) :

"I have very little to offer concerning an event perspective on linguistic events, and what I do have to say, I consider very tentative indeed".

III.3.2. Contraintes perceptives, contraintes articulatoires : quelques données

Nous allons illustrer la complexité des relations entre contraintes de production et contraintes de perception par deux exemples tirés de recherches menées sur l'organisation spatio-temporelle des gestes articulatoires dans notre laboratoire autour de C. Abry.

* Gestes audibles, gestes non audibles

Worley et Abry (1990 ; voir aussi Worley, 1989) ont analysé le mouvement de la lèvre supérieure pendant la production de séquences /iyiy/, puis /zizyzizy/ et /pipypipy/. Un des points importants de leur étude est qu'il existe dès le début de la voyelle /i/ un geste de protrusion vers le /y/, sans conséquence acoustique (pas de baisse du 3ème formant F3), ce geste étant suivi, le moment venu, par un geste rapide de protrusion qui a, lui, le résultat acoustique escompté (baisse de F3 d'une valeur proche de F4 pour /i/ vers une valeur proche de F2 pour /y/). Il existe ainsi des phases de préparation du geste, phases de geste non audible exploitant les non-linéarités du passage de l'articulatoire à l'acoustique (zones plateau dans la terminologie de Stevens), puis des phases de geste acoustiquement efficace, donc audible, et au timing assez précis (toujours pendant la consonne intervocalique, de manière à ne pas induire de perception de la voyelle /y/ avant cette consonne). Notons que ce geste non audible de prépration de la cible "exigeante" du /y/ (forte protrusion avec une aire aux lèvres extrêment réduite) est par cintre efficace visuellement (Escudier et al., 1989). Notons également, à ce propos, qu'un ensemble important de données sur l'intégration audio-visuelle en perception de parole suggèrent l'existence d'une "métrique conduit vocal" commune à ces deux modalités permettant de réaliser leur intégration avant la phase de décodage (Summerfield, 1987; Liberman et Mattingly, 1985), ce qui va bien avec la théorie motrice ou la théorie de l'action.

* Organisation temporelle et invariance dans la production d'oppositions de durée

Une littérature abondante existe en psychomotricité concernant l'existence dans la production de cycles moteurs d'un "patron de

phase" invariant (constance de la durée relative des phases du cycle par rapport à la durée totale du cycle, quelle que soit cette durée totale d'exécution). Divers travaux menés à l'ICP (Abry et al., 1988; Sock et al., 1988) montrent que s'il existe effectivement une invariance de certaines phases articulatoires ou acoustiques en fonction de la vitesse d'élocution, invariance qui pourrait être reliée à des contraintes générales des specifications des gestes dans les programmes moteurs, il existe également lorsque le débit varie des réorganisations qui permettent d'assurer des oppositions phonologiques associées à la durée telles que V vs V: (voyelle courte vs voyelle longue) ou C v CC (consonne simple vs consonne géminée).

III.3.3. Lindblom et "l'échelle hyper-hypo"

Bjorn Lindblom (1986-87, 1988), considérant successivement les chances de parvenir à définir une invariance articulatoire, acoustique ou perceptive et concluant dans chaque cas avec un certain scepticisme, remarque que l'ensemble des travaux menés en laboratoire concernent précisément une "parole de laboratoire" clairement articulée, et fait l'hypothèse que toute ébauche de découverte d'invariance dans un tel matériau "hyperarticulé" sera mise en défaut par des cas "écologiques" de matériaux "hypoarticulé" correspondant à la parole de tous les jours. Sa conclusion est qu'il ne faut pas chercher l'invariance dans le seul signal, mais dans un couple constitué de l'information contenue dans le signal et d'une information hors signal associée à l'ensemble des connaissances (ou des présupposés) qu'a l'auditeur concernant le signal émis par le locuteur. Ceci est décrit par le schéma proposé par Lindblom et représenté sur la Figure 2 : plus les attentes de l'auditeur (sa connaissance de la langue, du domaine, ...) sont pauvres, plus l'information contenue dans le signal doit être riche pour permettre une situation de communication réussie, d'où la nécessité pour le locuteur d'"hyperarticuler". Au contraire, pour un auditeur averti, le locuteur peut diminuer le débit d'information présent dans le signal (en "hypoarticulant") sans gêner la réussite de la communication.

Lindblom précise ainsi la notion d'invariance négociée (ou invariance adaptative, pour reprendre ses termes) que nous avons introduite au paragraphe II.3, en proposant finalement que seule une invariance combinant à la fois les niveaux articulatoire, acoustique, mais aussi linguistique, soit viable dans un cas général.

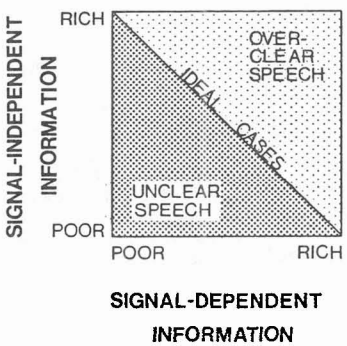

Figure 2 - Interaction locuteur-auditeur dans le cadre de la théorie adaptative (d'après Lindblom, 1986-87)

Conclusion : quelques propositions pour une stratégie de recherche

Nous l'avons vu dès l'introduction, tous les travaux sur le décodage acoustico-phonétique doivent proposer un "traitement" de la variabilité pour remonter au code, donc à un invariant symbolique. A ce titre, on peut dire qu'il n'est pas de théorie de la perception de parole qui ne soit aussi une théorie sur l'invariance. On peut tenter d'organiser les "stratégies vers l'invariance" autour de deux axes :

(i) apport plus ou moins important de "connaissances a priori" sur le signal ;

(ii) rôle plus ou moins grand des "bas niveaux" de traitement.

On peut ainsi dresser, par une combinatoire sur ces deux axes, un tableau de stratégies possibles :

1. Si l'on peut se passer de connaissances a priori, c'est que l'invariance existe dans le signal. Il "suffit" de la trouver, par les traitements mathématiques appropriés. Mais cette invariance peut être de bas ou de haut niveau.

Dans le cas d'une "invariance de bas niveau", il faut faire sur le signal les "bons calculs", les traitements correspondants pouvant être de pure algorithmique, ou guidés par les connaissances en physiologie ou en psychologie : c'est l'approche de Stevens.

Une invariance de "haut niveau" est une invariance statistique, calculée à partir de règles complexes sur le signal, règles qui relèvent en général des formalismes de systèmes experts : "si j'ai reconnu préalablement tel phonème avec telle probabilité, et que je mesure maintenant tel paramètre à telle valeur, alors je suis sans doute sur une portion de signal correspondant à telle classe".

2/ Si l'on considère que le signal acoustique ne comporte pas toute l'information nécessaire pour le décodage, alors il faut essayer de combler ce déficit d'information, et ceci peut là encore se faire à haut ou bas niveau.

Les contraintes de bas niveau en parole sont typiquement celles définies par Poggio pour la vision, celles qui permettent de "régulariser" le problème de l'inversion, afin de récupérer une invariance quelque part au niveau moteur à partir du signal acoustique. Ces contraintes seraient alors sans doute définies sur le mouvement des articulateurs (contraintes mécaniques, contraintes d'économie).

Les contraintes de haut niveau sont, pour la parole, les plus classiquement utilisées : elles viennent de prédictions ou de vérifications faites par les niveaux supérieurs de traitement, lexical, syntaxique, sémantique, pragmatique, et permettent de "contraindre" la recherche du code en fonction des résultats antérieurs de cette recherche. L'invariant n'existe alors qu'au niveau phonologique.

Il est clair que l'accent a été mis dans ce texte sur les niveaux d'explication bas au détriment des niveaux hauts. Je me suis résolument mis du côté d'une perspective néo-réaliste, à la Marr/Stevens ou à la Gibson-Poggio/Liberman-Fowler. Deux points me semblent importants à signaler en conclusion pour décrire notre perspective de recherche à l'ICP, recherche dont j'ai donné quelques échos.

1. Le choix, contraire au néo-réalisme, d'une approche cognitiviste, descendante, approche de type extraction d'indices acoustiques quantitatifs, contextuels, formant un ensemble partiellement contradictoire, ensemble sur lequel un niveau supérieur de traitement doit exercer un raisonnement pour parvenir à une catégorisation, est un choix parfaitement défendable du point de vue des données existant actuellement dans le domaine. Si nous penchons dans notre laboratoire plutôt pour une approche néo-réaliste, ce n'est pas tant à cause d'une conception générale des mécanismes de perception humains, que pour une raison stratégique, d'efficacité scientifique. Il nous semble en effet que la perspective néo-réaliste a le mérite de poser clairement les problèmes et les enjeux, en évitant le piège d'une perspective néo-idéaliste dans laquelle tout est possible au niveau acoustique, puisque tout se règlera dans des mécanismes complexes au niveau supérieur, mécanismes qu'il est en général difficile de définir et même de contraindre suffisamment.

2. Par contre, dans ce cadre néo-réaliste, le choix d'une stratégie du "tout articulatoire" ou d'une stratégie du "tout acoustique" (ou du "tout perceptif") nous semble difficile à tenir. Nous leur préférons une recherche des contraintes tant du système de production que du système de perception, tout en étant conscients que l'étape décisive à franchir dans les prochaines années sera celle de la description précise des conditions de négociation entre le locuteur et l'auditeur, précisant pour chacun d'entre eux ce qui peut être négocié (et sous quelle forme) et ce qui ne le peut pas.

Remerciements

Ce texte a été amélioré par les critiques de Pierre Escudier. Il s'appuie d'ailleurs sur des années de discussions et de recherches menées en commun à l'ICP. Il s'appuie également sur un certain nombre de pistes de réflexion ouvertes par Christian Abry. Que l'un et l'autre en soient très sincèrement remerciés.

Bibliographie

Abry C, Boë LJ, Schwartz JL (1989), J. Phonetics 17, 47-54
Abry C, Schwartz JL (1987), Bull. LCP 1A, 191-210
Abry C, Orliaguet JP, Sock R (1988), Workshop LASCO3, 133-166
Assman PF, Nearey TM, Hogan JT (1982), JASA 71, 975-989
Beautemps D.(1989), DEA, INPG, Grenoble.
Blumstein SE (1983), in Invariance and Variability in Speech Processes, JS Perkell & D Klatt eds., Lawrence Erlbaum Associates
Blumstein SE, Stevens KN.(1979), JASA 66, 1001-1017
Blumstein SE, Stevens KN, GN Nigro (1977), JASA 61, 1301-1313
Chomsky N, Halle M (1968), The Sound Pattern of English, Harper-Row
Delgutte B.(1980), JASA 68, 843-857
Delgutte B (1982) in The Representation of Speech in the Peripheral Auditory System, R Carlson & B Granström eds., Elsevier Biomedical
Delgutte B (1983) in Invariance and Variability in Speech Processes, JS Perkell & D Klatt eds., Lawrence Erlbaum Associates, 163-177
Diehl RL, Buchwald McCusker S, Chapman LS (1981) JASA 69, 239-248
Escudier P, Benoît C, Lallouache T (1990), 1er CFA, SFA
Fabre JP(1988), La Recherche, 202, 1104--1106
Fowler CA (1983), J. Phonetics 11, 303-322
Fowler CA (1986), J. Phonetics 14, 3-28
Fowler CA, Rubin P, Remez RE, Turvey MT.(1979), in Language production, B Butterworth ed., New-York : Academic Press
Fowler CA, Shankweiler DP (1978), JASA 63, S4(A)
Gibson JJ (1966), The Senses Considered As Perceptual Systems, New-York, Boston : Houghton Mifflin
Gottfried TL (1979), Speech Communication Papers, 97th ASA Meeting, JJ Wolf & DH Klatt eds., ASA, 29-32
Hoffman D (1984), in La Perception Visuelle, C Bonnet ed., Bibliothè que Pour la Science, Diffusion Belin, Paris, 110-116
Horn BKP.(1975), in The psychology of Computer Vision, PH Winston ed., McGraw-Hill Publ., New-York, 115-155
Jakobson R, Fant G, Halle M.(1963), Preliminaries to Speech Analysis, Cambridge, MA:MIT Press
Johansson G (1950), Configurations in Event Perception, Uppsala : Almquist and Wiksell.
Johansson G (1964), in Perception : Essays in honor of James J Gibson, R MacLeod & H Pick eds., Ithaca : Cornell University Press
Johansson G (1984), Computational Models of Hearing and Vision, Tallinn, 100-102
Kewley-Port D (1980), Res. Speech Percept. Tech. Rep. 3, Indiana Univ.
Kewley-Port D (1983), JASA 73, 322-335
Liberman AM, Cooper FS, Shankweiler D, Studdert--Kennedy M (1967), Psychological Rev., 74, 431-461
Liberman AM, Mattingly IG (1985), Cognition, 21, 1-36
Lindblom B (1963), JASA 35, 1773-1781
Lindblom B (1968), Doct. Thesis, University of Lund
Lindblom B (1986-87), Perilus Institute of Linguistics, Stockholm, 2-20
Lindblom B (1988), in Working Models of Human Perception, Academic Press, London
Macchi M (1980), JASA 68, 1636-1642
Marr D (1982), Vision, W.H. Freeman and Company, San Francisco
Marr D, Poggio T. (1979), Proc. R. Soc. Lond. B. 204, 301-328
Perrier P, Abry C, Keller E (1989), J. Acoustique 2, 69-77
Petitot J (1985), Les Catastrophes de la Parole de Roman Jakobson à René Thom, Maloine, Paris
Pisoni DB, Carrell TD, Simnick SS (1979), Speech Comm. Papers, 97th ASA Meeting, J.J. Wolf & D.H. Klatt eds., ASA, 19-24
Poggio T.(1984), Computational Models of Hearing and Vision, Tallinn, 123-127
Schaeffer P (1966), "Traité des objets musicaux", Editions du Seuil, Paris
Schwartz J.L (1987), Bull. LCP. 1A, 159-190
Schwartz JL, Boë LJ, Perrier P, Guérin B, Escudier P (1989), Eurospeech Paris, 63-66.
Schwartz JL, Escudier P (1987), in The Psychophysics of Speech Perception, MEH Schouten ed., NATO-ASI Series, Martinus Nijhoff.
Schwartz JL, Escudier P (1979), Speech Comm. 8, 235-259
Sock R, Ollila L, Delattre C, Zilliox C, Zohair L (1988), J. Acoustique 1, 339-345.
Stevens KN (1972), in Human Communication : A Unified View, EE Davis & PBDenes eds., New-York : McGraw Hill, 51-66
Stevens KN (1980), JASA 68, 836-842
Stevens KN (1989), J. Phonetics 17, 3-46
Stevens KN, Blumstein SE (1978), JASA 64, 1358-1368.
Strange W, Verbrugge RR, Shankweiler DP, Edman TR (1976), JASA 60, 213-224.
Summerfield Q (1987), in Hearing by Eye : The Psychology of Lipreading, B Dodd & R Campbell eds., Lawrence Erlbaum Associates.
Terzopoulos D.(1984), PhD Thesis, Dept. of Electrical Engineering and Computer Science, MIT
Torre V, Poggio T (1984), MIT Artificial Intelligence Laboratory Memo N°776
Ullman S (1979), Proc. R. Soc. Lond. B. 203, 405-426
Verbrugge RR, Shankweiler DP, Fowler CA (1979) Speech Comm. Papers, 97th ASA Meeting, JJ Wolf & DH Klatt eds. ASA, 15-18
Worley C (1989), Thèse de l'INPG, Grenoble.
Worley C, Abry C (1990), 1er CFA, SFA.
Wu ZL, Escudier P, Schwartz JL, Sock R (1988), 17èmes JEP, SFA, 219-224.
Wu ZL, Escudier P, Schwartz JL (1989), IEEE-ICASSP, 2013-2016