La parole multisensorielle : Plaidoyer, problèmes, perspective

Jean-Luc Schwartz

Institut de la Communication Parlée – CNRS UMR 5009
Institut National Polytechnique de Grenoble, Université Stendhal
Mél : schwartz@icp.inpg.fr

ABSTRACT

Since more than 20 years, ICP has set multisensoriality as a key component of its research agenda. More and more, it appears that the multisensorial nature of speech is neither arguable, nor marginal, and that it belongs to the very core of speech communication. Moreover, it has renewed or enhanced speech research in many domains, and particularly: processing, representations, invariance, perceptuo-motor links, cortical circuits, adaptability, development, handicaps, learning and technology. After a general presentation (a «plaidoyer») on the importance of multisensoriality in speech communication, we focus on «problems» about audio-visual fusion, including architectures, fusion control, and very early interactions. We discuss model comparison in the Bayesian framework, and present recent data on speech enhancement exploiting novel audio-visual techniques. We conclude by a «perspective» on some hot topics in the study of audio-visual speech perception and perceptuo-motor links.

INTRODUCTION

Depuis plus de 20 ans, l'ICP a fait de la multisensorialité de la communication parlée un élément central de sa recherche. Cette position n'a pas toujours été bien comprise. Alors, quoi ? Le son ne suffit-il pas à se faire entièrement et parfaitement comprendre ? Et ne pose-t-il pas assez de problèmes comme ça ? La multisensorialité de la parole a été parfois considérée comme, au mieux, un luxe, au pire, un caprice, pour chercheurs peu soucieux de se confronter à la vraie nature du langage, sonore, bien sûr !

Et puis, de plus en plus, il apparaît que la multisensorialité de la parole n'est ni contestable, ni marginale. D'abord parce qu'elle est au cœur du dispositif de la communication orale : c'est ce que je rappellerai dans mon plaidoyer initial. Ensuite et surtout, parce qu'elle oblige à se poser des questions sur tout ce qui importe dans le domaine, et particulièrement : sur les traitements, sur les représentations, sur les circuits corticaux, sur les relations perception-action, sur l'adaptabilité, sur le développement, sur l'apprentissage, sur les handicaps, sur l'émergence, sur les technologies. De ces questions multiples, je reprendrai deux ou trois « problèmes » concernant une question particulière sur laquelle je travaille depuis longtemps, celle de la nature du processus de fusion audio-visuelle. Je les reprendrai sous une forme que j'ai déjà pratiquée dans des articles précédents [SCH98, SCH02] (où l'on trouvera une grande partie des références utiles pour cet article), mais avec une série de résultats nouveaux et significatifs, qui sont venus conforter ou préciser les propositions que nous faisions depuis un certain nombre d'années. Et puis, je tenterai une échappée plus générale, pour tracer, donc, une perspective, ou plutôt une mise en perspective des recherches dans un tableau dans lequel la multisensorialité est au cœur du projet.

PLAIDOYER

Donc, la parole, qu'on le veuille ou non, est multisensorielle. Cette multisensorialité s'impose à notre système de perception, irrésistiblement (et même on le verra, irrépressiblement). Elle s'impose au chercheur curieux, non moins irrésistiblement ! Par ondes concentriques, du plus périphérique (ou marginal), au plus central (et essentiel), rappelons l'exposé des motifs.

Le conduit vocal peut être entendu, vu ... et touché

Le célèbre psychologue Robert H. Stetson écrivait dès les années 20 dans *Motor Phonetics* : «Plutôt qu'une série de sons produits par du mouvement, la parole est une série de mouvements rendus audibles». Ces mouvements audibles sont aussi visibles, on le sait bien : c'est la lecture labiale, bien utile aux sourds ... et aux autres, et qui permet d'appréhender environ 40 à 60 % des phonèmes d'une langue donnée, et de 10 à 20 % des mots – bien que la variabilité interindividuelle à ce niveau soit considérable, les meilleurs lecteurs labiaux (généralement des sourds) atteignant des scores supérieurs à 60 %. Ils peuvent également être touchés avec profit : c'est la méthode Tadoma, qui permet aux sourds aveugles, en plaçant le pouce sur les lèvres du locuteur, l'index sur la joue et les autres doigts sur le cou en dessous de la mandibule, d'atteindre des scores d'identification phonétique de 40 à 60 % comparables à la lecture labiale, bien que le type de confusions soit très différent.

Chacun sait lire sur les lèvres

Si la lecture labiale est connue pour fournir une aide précieuse aux sourds et malentendants, chacun de nous sait lire sur les lèvres, et s'en sert dans la vie quotidienne. C'est particulièrement vrai dès qu'il y a du bruit dans le système, comme le montrent les fameuses expériences de Sumby et Pollack. Mais c'est également le cas en l'absence de bruit, lorsque la tâche de compréhension est complexe, et particulièrement en langue étrangère.

Et on ne peut pas s'en empêcher !

C'est la leçon de l'effet McGurk, maintenant bien connu, dans lequel le montage d'un [b] auditif superposé à un [g] visuel conduit à la perception d'un [d] (ou, en anglais, d'un [th]). Quoique, à vrai dire … on parvient à s'en empêcher, lorsque le conflit est trop important, comme dans le cas de films doublés, par exemple (on trouvera, à ce sujet, d'intéressantes données sur les interactions entre cohérence audiovisuelle et masquage, dans [BRU03]). Cependant, chassez la multimodalité, elle revient au galop : car c'est alors un autre système de traitement de l'information, celui de la localisation, qui subit l'influence d'interactions audiovisuelles, et le mouvement des lèvres, s'il ne parvient plus à peser efficacement sur le système de compréhension, attire irrésistiblement à lui la localisation de la source sonore. C'est ainsi que fonctionne l'effet ventriloque.

L'enfance multisensorielle

Cette multisensorialité de la parole, l'enfant y baigne dès sa naissance. Alors, tout naturellement, ses effets se révèlent dans le développement du langage. Ainsi, la préférence marquée pour les bilabiales au début du babillage dans les premiers stades de l'acquisition du langage est renforcée chez les enfants malentendants, mais diminue chez les enfants aveugles : et voilà, madame, pourquoi votre enfant vous appelle « maman », « mummy », « Mutti » ou « mamma ». L'absence d'entrée visuelle peut d'ailleurs se traduire par des difficultés à acquérir certains contrastes phonétiques, et notamment la distinction entre [m] et [n], peu audible mais bien visible. Et les travaux de Leybaert et coll. montrent que la modalité visuelle peut permettre d'acquérir parfaitement la phonologie de sa langue (à défaut, bien sûr, d'une pleine maîtrise du contrôle orofacial) grâce au langage parlé complété, qui consiste à adjoindre aux gestes labiaux des gestes manuels synchronisés, levant les ambiguïtés de la lecture labiale.

De l'ontogenèse à la phylogenèse

Ainsi, tout naturellement, on en vient à supposer que, étendant la formule de Jakobson, « you speak to be heard … *and seen* in order to be understood ». Pour s'en convaincre, il suffit de se rappeler de ces moments où, lors d'un cours ou d'une conférence un peu fastidieuse, on cherche à faire passer, aussi silencieusement que possible, un message à un voisin un peu éloigné. Spontanément, on se met à murmurer, en exagérant le plus possible le mouvement des lèvres … pour être mieux vu ! Et ce qui est extrêmement frappant dans cette expérience, si l'on y songe, c'est que l'on sache faire cela, sans effort, sans difficulté et sans apprentissage.

Bref, on sait moduler au choix l'un ou l'autre des canaux de sortie, parler fort ou doucement, mais aussi, parler « visiblement » ou « invisiblement », selon que l'on veut être vu, ou au contraire le moins vu possible : voir [BEA99], pour une étude des effets de l'hyperarticulation visuelle sur la perception des voyelles et des consonnes. C'est d'ailleurs sans doute à cause de sa bonne visibilité que le contraste [m] vs. [n] existe dans presque toutes les langues du monde (94% des 317 langues de la base de données UPSID, Boë et Vallée, communication personnelle), car son faible contraste auditif l'aurait sans doute peu favorisé, dans le cadre des théories de l'émergence des systèmes phonologiques « à la Lindblom ».

Conclusion

Nous voilà bien au cœur du processus, comme promis. Au cœur d'un processus de communication orale dans lequel la vision est présente depuis toujours, et (presque) en toutes circonstances. Et maintenant, au centre d'un ensemble de recherches qui ont vu la multisensorialité s'attaquer successivement à (presque) tous les domaines d'étude : l'attention et la mémoire, la détection de parole, l'analyse de scènes, la phonétique, les interférences culturelles, l'apprentissage, le développement, les théories de l'émergence, la prosodie, les émotions, etc. Mais, plus encore, la multisensorialité s'impose, parce qu'elle nous force à embrasser le système de communication dans sa globalité, et à poser différemment les questions fondamentales de la perception, de l'action, et de l'interaction. C'est ce que je tenterai de montrer en perspective.

PROBLEMES

Je vais maintenant centrer le propos sur la question des mécanismes de fusion audiovisuelle pour la perception et la reconnaissance de la parole. L'approche que nous avons suivie dans ce domaine depuis 15 ans croise des hypothèses sur les architectures et les mécanismes de fusion, avec des mises en œuvre de modèles quantitatifs et de procédure de comparaison et d'évaluation, à l'aune de données de psychologie expérimentale, originales ou piochées dans la littérature.

Architectures de fusion

Les données comparatives que nous avons passées en revue montrent que les voies auditive et visuelle convergent –en un lieu qu'il s'agit de déterminer, anatomiquement et fonctionnellement– de sorte que la prise de décision (phonétique et/ou lexicale) reflète à la fois les deux types d'information. Notamment, les données tant sur l'effet McGurk que sur la perception en milieu bruité illustrent bien l'existence d'interactions, dont il s'agit de rendre compte dans une modélisation du/des processus de fusion audiovisuelle. Notons que ce problème de fusion de capteurs intéresse à la fois modélisateurs (et psychologues) et technologues, puisqu'il s'agit aussi de déterminer comment combiner son et image de manière aussi efficace que possible pour construire un système de reconnaissance de parole audiovisuelle robuste dans le bruit.

Avec Jordi Robert-Ribes, nous avons longuement travaillé sur les architectures de fusion. Nous avons pu montrer, en croisant modèles généraux de psychologie et fusion de capteurs, qu'il existait essentiellement 4 architectures possibles, rappelées dans la Fig. 1 (d'après [SCH98]). Le modèle à « Identification Directe » (ID, Fig. 1a) compile les deux flux de données et effectue directement l'étape de classification sur cette entrée multisensorielle, sans étape intermédiaire de mise en forme commune des données. Dans le modèle à « Identification Séparée » (IS, Fig. 1b), il y a classification phonétique séparée du son et de l'image du locuteur, suivie d'un processus de fusion que l'on appelle tardive, car elle suit l'accès au code dans chaque modalité. La classification peut concerner soit l'ensemble des traits phonétiques, soit seulement certains traits supposés spécifiques à chaque modalité. La fusion est ensuite une fusion de décisions, qui est en général effectuée par un processus probabiliste fourni par la théorie de la décision.

Dans les architectures de fusion dites précoces, l'intégration précède l'accès au code. Mais il reste à spécifier la forme que peut prendre le flux commun de données après fusion. On peut d'abord estimer que l'audition est pour la parole la modalité dominante et que l'entrée visuelle est recodée sous un format compatible avec celui des représentations auditives : c'est le modèle de Recodage dans la modalité Dominante (RD, Fig. 1c). On peut supposer, par exemple, qu'audition et vision convergent vers un processus d'estimation de la fonction de transfert du conduit vocal, les caractéristiques de source étant réservées à la seule entrée auditive, et la décision s'effectuant ensuite sur ces deux informations partielles. On peut aussi s'inscrire dans le cadre d'une théorie faisant jouer un rôle crucial à la cause commune du son et de l'image –qu'il s'agisse de la Théorie Motrice ou de la Théorie de la Perception Directe– et supposer qu'il existe un Recodage commun des deux entrées sensorielles vers la modalité Motrice (RM, Fig. 1d). Les caractéristiques principales des gestes articulatoires, estimées conjointement sur les modalités auditive et visuelle, sont ensuite fournies à un processus de classification qui permet l'accès au code.

Cette taxonomie, adaptée et modifiée de celle de Summerfield [SUM87], présente l'intérêt de prendre en compte les théories classiques des interactions entre modalité en

psychologie expérimentale, mais aussi les principes de fusion de capteurs en théorie de la décision. Elle a été souvent reprise, à la fois par les psychologues et par les spécialistes de reconnaissance de la parole audiovisuelle. Parmi ces 4 architectures, les modèles ID et IS sont sans conteste les plus utilisés en reconnaissance de parole (voir une présentation d'un large ensemble d'architectures de décodage dans [SCH02] ; et des campagnes de comparaison d'architectures dans [ROB95] ; [TEI99]). Des arguments tirés de données de psychologie expérimentale nous ont plutôt conduit à privilégier le modèle RM : nous y reviendrons dans la partie Perspective.

Contrôle du processus de fusion

Fusion en traitement de l'information

Une fois fait le choix d'une architecture de fusion, le spécialiste de traitement de l'information sait qu'il lui faut résoudre un second problème, d'une importance capitale pour l'efficacité de son système de reconnaissance de parole par exemple : celui de la nature du mécanisme de fusion, et plus particulièrement de son contrôle en fonction d'un certain nombre de paramètres. La taxonomie proposée par Isabelle Bloch [BLO96] est à cet égard claire et utile. Elle distingue trois types de processus qui diffèrent selon leurs comportements. Le plus simple est le processus Indépendant du Contexte et à Comportement Constant (ICCC), qui opère sur les entrées une fusion dont la loi est fixe (par exemple, additive ou multiplicative). On peut aussi considérer un processus Indépendant du Contexte et à Comportement Variable (ICCV), dont la loi de fusion est variable selon les valeurs des entrées. On peut imaginer ainsi qu'un système effectue une fusion additive avec des coefficients de pondération différents suivant le niveau des entrées, permettant par exemple de pondérer plus l'entrée la plus importante. Enfin, un processus Dépendant du Contexte (DC) prend en compte des connaissances sur l'environnement extérieur, qui peuvent lui permettre par exemple de pondérer plus ou moins une entrée selon la nature de ces informations contextuelles. Pour la reconnaissance de la parole, le cas le plus classique est celui de l'utilisation d'un processus dépendant du contexte, dans lequel une variable estimant le niveau de bruit acoustique (et/ou visuel, dans certaines applications) vient réguler l'importance relative des flux auditif et visuel. Cette variable peut être estimée soit directement sur le signal, soit en sortie de classifieur, par exemple dans le cas du modèle IS. On trouvera dans [TEI99] une illustration de l'implémentation de fusion DC pour les 4 architectures de la Fig. 1.

Fusion en perception de parole

Si l'on en revient au domaine de la psychologie expérimentale, les données suggèrent l'existence d'un certain nombre de contrôles possibles. Il faut d'abord prendre en compte la variabilité inter-individuelle. En effet, les sujets accordent plus ou moins d'importance à la modalité visuelle dans le processus d'intégration. Ainsi, Cathiard [CAT 94] montre dans des expériences de conflit audiovisuel l'existence de deux groupes, différant selon le poids qu'ils accordent à l'entrée visuelle dans la gestion perceptive du conflit. Les caractéristiques de la perception audiovisuelle semblent également dépendre d'influences culturelles, ainsi que le suggèrent les travaux menés par Sekiyama qui montrent que la communauté japonaise attribue apparemment moins de poids à la modalité visuelle dans des expériences de conflit [SEK 93]. Ce thème a concentré depuis quelques années une série de travaux expérimentaux et d'interprétations contradictoires : marque d'une communauté qui n'est pas habituée à regarder le visage du locuteur, ou simple conséquence de variations dans l'inventaire des prototypes linguistiques, voire même effet d'une langue à tons qui inciterait ses utilisateurs à se concentrer plus sur la modalité auditive. Enfin, des facteurs attentionnels semblent capables de modifier les poids de l'une ou l'autre modalité, bien que l'effet Mc Gurk dresse une borne à ces facteurs, puisque cet effet est cognitivement impénétrable, et résiste à l'orientation sélective de l'attention vers la composante auditive du stimulus conflictuel.

Fusion contrôlée vs. constante : le modèle FLMP

Bien que les données semblent plaider en faveur d'un processus de fusion dépendant du contexte, le fait n'est pas réellement acquis expérimentalement. En effet, si l'on se livre à un sondage sur *le* modèle par excellence dans le domaine de la parole audiovisuelle, un nom risque de revenir en force : celui du FLMP, « Fuzzy Logical Model of Perception », de Dominic Massaro [MAS98]. Ce modèle de perception à logique floue est d'ailleurs connu très au-delà du domaine de la perception de la parole, multisensorielle ou non. Le principe en est simple, c'est celui d'une architecture IS (estimation séparée des classes phonétiques dans chaque modalité) suivie d'une fusion ICCC multiplicative. Ce modèle a été appliqué par Massaro à de multiples problèmes (identification auditive de traits phonétiques, perception audiovisuelle des émotions, reconnaissance de caractères, identification de la face, etc).

Dans le cas qui nous occupe ici, le principe du modèle est simple. Si l'on considère une expérience d'identification audiovisuelle parmi N classes possibles, audition et vision estiment chacune la vraisemblance monosensorielle de chaque catégorie (ce que Massaro nomme le « degree of auditory or visual support »), soit a_i et v_i. Puis la probabilité de sélectionner la classe i est estimée, en modalité audiovisuelle, par la loi :

$$p_{AV}(C_i) = \frac{a_i v_i}{\sum_j a_j v_j}$$

Il s'agit alors de déterminer, par une procédure d'optimisation, les valeurs des paramètres a_i et v_i permettant d'approcher au mieux les données expérimentales d'identification auditive, visuelle et audiovisuelle. Le principe du FLMP est donc clairement différent d'un processus de fusion dépendant du contexte. Or, le modèle FLMP s'avère capable de s'ajuster à la plupart des données expérimentales disponibles, même lorsque celles-ci semblent suggérer fortement l'existence de variables contextuelles contrôlant la fusion : voir par exemple le rejet par Massaro [MAS93] de différences inter-linguistiques, et les difficultés rencontrées par Tiippana et al. [TII01] pour interpréter leurs données, avec des variations considérables des réponses des sujets à des stimuli audiovisuels selon l'attention visuelle (avec moins de « poids » apparent de l'entrée visuelle lorsque l'attention est distraite), et cependant un excellent fit de ces données par le modèle FLMP ! Où est l'erreur ?

Une critique méthodologique du modèle FLMP

Le petit bout d'histoire que je vais raconter maintenant me semble important, d'abord parce que le débat autour du FLMP a souvent obscurci les raisonnements, et nous avons nous-mêmes mis (trop) longtemps à y voir un peu plus clair ; mais aussi parce que le développement qui suit a une portée générale sur la méthodologie de comparaison de modèles phénoménologiques.

Les travaux sur le FLMP se sont focalisés sur un index unique : l'erreur quadratique moyenne sur les réponses des sujets sur les N classes, sommées sur toutes les conditions expérimentales disponibles (auditives, visuelles, audiovisuelles conflictuelles ou non), en comparant ces réponses aux prédictions du modèle :

$$rmse = \sqrt{\sum_{i,E}(R_E(C_i) - p_E(C_i))^2}$$

où $R_E(C_i)$ et $P_E(C_i)$ sont respectivement les réponses des sujets et les probabilités calculées par le modèle pour la condition expérimentale E, et la classe C_i. Or, dans le cas de stimuli conflictuels tels que ceux utilisés souvent dans la comparaison de modèles (et notamment sur l'effet McGurk), on risque de rencontrer des situations où les stimuli auditifs et visuels combinés produisent des réponses $R_A(C_i)$ et $R_V(C_i)$ quasiment incompatibles, c'est-à-dire qu'une modalité répond « blanc » pendant que l'autre répond « noir » (ainsi, dans l'effet McGurk , la modalité auditive répond [b] pendant que la modalité visuelle répond [g]). Le problème est que, dans ce cas, le modèle FLMP risque de combiner des valeurs a_i et v_i telles que :

$$\forall i \quad a_i v_i = 0$$

En conséquence :

$$p_{AV}(C_i) = \frac{a_i v_i}{\sum_j a_j v_j} = \frac{0}{0} \qquad \textit{indéterminé}$$

On peut aisément montrer que, dans ce contexte, le modèle FLMP est effectivement capable de prédire n'importe quel jeu de valeurs $R_{AV}(C_i)$, et donc de s'ajuster à toutes les données possibles [SCH03]. Là encore, où est l'erreur ?

Le formalisme bayésien

Chercher à démontrer qu'un modèle phénoménologique s'ajuste bien aux données est classique, notamment en psychologie expérimentale. Cependant, le modèle peut en général être au départ exprimé sous forme probabiliste, caractérisant la probabilité que les données expérimentales aient été produites par le modèle. Cette probabilité *a priori*, $p(D/M)$, est proportionnelle à la probabilité *a posteriori* $p(M/D)$ qui exprime la probabilité que le modèle M soit correct, connaissant les données D. Si le modèle M dépend de paramètres π, on cherche en général à estimer le « meilleur fit » du modèle , soit :

$$\pi_0 = \arg\max_\pi(p(D/M(\pi))$$

Si le modèle $M(\pi)$ fournit pour chaque composante D_i des données une modélisation gaussienne de moyenne m_i et de variance constante σ^2, alors on a simplement :

$$\log(p(D/M(\pi))) = \sum_i \log(p(D_i/M(\pi))) = cte - {}^1\!/_{2\sigma^2} \sum_i (m_i(\pi) - D_i)^2$$

et on obtient :

$$\pi_0 = \arg\min_\pi(rmse(\pi))$$

Ainsi, la minimisation d'une erreur quadratique moyenne de prédiction des réponses est bien ancrée sur un calcul de maximum de vraisemblance (en réalité, dans le cas de données de classification, on traite de lois binomiales ou polynomiales et non de lois gaussiennes, et donc la maximisation de la vraisemblance ne correspond plus à la minimisation de l'erreur de prédiction ; mais elle s'en éloigne assez peu en général).

OUI MAIS … Jaynes, le théoricien de la modélisation et du raisonnement bayésien [JAY] nous rappelle un élément crucial de la comparaison de modèles. Si vous cherchez à comparer deux modèles M_1 et M_2, caractérisés chacun par un jeu de paramètre π_1 et π_2, ce qui doit intervenir dans la comparaison, ce n'est pas leur « meilleur profil », c'est-à-dire la comparaison des vraisemblances maximisées : $p(M_1(\pi_{01})/D)$ et $p(M_2(\pi_{02})/D)$ où π_{01} et π_{02} sont respectivement les jeux de paramètres maximisant les vraisemblances de M_1 et M_2 connaissant les données, mais bien leur vraisemblance *totale*, soit $p(M_1/D)$ *et* $p(M_2/D)$. En passant des probabilités *a posteriori* aux probabilités *a priori* par la formule de Bayes, les termes à comparer sont de la forme:

$$p(D/M_i) = \sum_{\pi_i} p(D/M_i(\pi_i))$$

Et c'est là que tout va se jouer. Si un modèle, comme le FLMP, est capable de « tout prédire », la sanction, en termes bayésiens, est immédiate. Certes, il existe pour chaque jeu de données Δ (les « bonnes » données, D, comme les « mauvaises »), un jeu de paramètres π_Δ permettant de bien ajuster le modèle FLMP aux données, soit :

$$\pi_\Delta = \arg\max_\pi(p(\Delta/FLMP(\pi)) \quad et \quad p(\Delta/FLMP(\pi_\Delta)) \; élevé$$

mais, comme toutes les prédictions sont possibles avec une égale qualité, pour les « bonnes » données D, l'espace de paramètres π_D acceptable en est réduit d'autant, puisqu'il faut bien partager l'espace des paramètres disponibles entre toutes les prédictions possibles, les bonnes comme les mauvaises. En conséquence, la probabilité globale du modèle FLMP est faible :

$$p(D/FLMP) \quad faible$$

Cette histoire de modélisation a une morale, bien sûr : si un modèle peut prédire tout, le vrai comme le faux, avec un égal bonheur (ce qui est, on le sent bien, peu acceptable : immoral, pour un modèle !), alors sa vraisemblance totale par rapport aux données expérimentales disponibles sera faible.

Prenons un exemple simple. Supposons que l'on compare deux modèles, soit FLMP, et un modèle concurrent que nous appellerons « Auditif », reposant sur l'hypothèse selon laquelle seule la modalité auditive intervient dans la fusion audiovisuelle. Pour ce modèle AUD :

$$p_{AV}(D/AUD) = p_A(D/AUD)$$

Considérons un problème à deux catégories, avec des données expérimentales fictives telles que la probabilité de réponse dans la catégorie 1 soit respectivement 0.99 en condition auditive, 0.01 en condition visuelle et 0.95 en condition audiovisuelle. Le modèle AUD fournit un ajustement acceptable de ces données, avec une erreur de prédiction $rmse_{AUD} = 0.0283$, ce qui signifie que les valeurs expérimentales sont ajustées à moins de 0.03 près. Le modèle FLMP, fort de sa capacité à s'ajuster à toute situation conflictuelle, donne un bien meilleur ajustement, soit $rmse_{FLMP} = 0.0122$. Et pourtant, on sent bien que le modèle AUD devrait sortir renforcé par ces données fictives, qui semblent effectivement indiquer que la réponse audiovisuelle est peu influencée par l'entrée visuelle. Précisément, le calcul des probabilités globales confirme cette analyse. Si l'on teste, pour les deux modèles AUD et FLMP, tous les jeux possibles de paramètres $\pi=(P_A(C_1),P_V(C_1))$, on constate effectivement que la valeur de π permettant au modèle FLMP un bon ajustement aux données doit être choisie extrêmement précisément, et que l'erreur d'ajustement s'accroît dramatiquement dès que l'on s'écarte de cette valeur optimale ; tandis que pour le modèle

AUD, la gamme de valeurs acceptables est plus importante. En conséquence, si l'on estime les probabilités globales en intégrant les probabilités de chaque modèle sur l'espace des paramètres, par échantillonnage ou par une méthode de Monte-Carlo [SCH03], on obtient - une vraisemblance du modèle AUD près de 5 fois plus grande que celle du modèle FLMP. Ouf !

En conclusion …

Le message est simple : au calcul d'une erreur d'ajustement d'un modèle, préférez toujours, lorsque vous le pouvez, le calcul d'une vraisemblance globale, certes beaucoup plus lourde, mais seule capable de démontrer que non seulement votre modèle peut prédire les données réelles, mais encore qu'il peut aussi … ne pas prédire les données fausses ! On peut ainsi, dans ce cadre bayésien, montrer qu'effectivement, la fusion audiovisuelle n'est pas constante, mais dépend notamment du sujet –et vraisemblablement, comme indiqué précédemment, de l'attention, du niveau de bruit, voire de la langue. C'est donc bien un processus de fusion dépendant du contexte qui semble à l'œuvre dans la perception audiovisuelle de la parole.

Des interactions très précoces

Nous voilà dotés d'une taxonomie d'architectures de fusion –sur laquelle nous reviendrons dans notre « Perspective »– et de solides raisons de penser que cette fusion dépend d'ingrédients contextuels. Tout au long de cette section, nous avons admis que la fusion opérait *postérieurement à la prise d'information*, qui restait, elle, monosensorielle : dans les architectures de la Fig. 1, il y a d'abord extraction indépendante des paramètres auditifs et visuels, puis fusion. Or il y a une donnée majeure qui émerge de la littérature de ces 10 dernières années, et c'est précisément la mise en évidence de mécanismes d'interaction très précoces, intervenant au cœur même des processus d'extraction d'information auditive et visuelle. C'est l'histoire de cette émergence que nous allons raconter maintenant.

De Driver à Grant, des mises en évidence indirectes

Montrer l'existence de mécanismes précoces est délicat, car ils sont de prime abord difficiles à séparer de mécanismes plus classiques de lecture labiale. La première belle mise en évidence est l'œuvre de Jon Driver, spécialiste d'interactions sensori-motrices. Dans un article paru dans Nature en 1996 [DRI96], il propose une série d'expériences astucieuses, dont la plus simple est la suivante. Dans un mixage de deux voix acoustiquement superposées, il apparaît que voir un des locuteurs permet de mieux comprendre … l'autre. Or, on ne peut pas plaider ici que la vision intervient par la lecture labiale, puisqu'on ne peut pas lire les bonnes lèvres. Il faut plutôt plaider, comme le fait Driver, un mécanisme « d'attention sélective multisensorielle », ou, comme nous l'avons nous-mêmes réinterprété, l'existence d'un système d'Analyse de Scènes de parole Audio-Visuelle (ASAV) généralisant à un ensemble de signaux de modalités multiples les principes de cohérence, fusion, fission proposés par Bregman dans l'Analyse de Scènes Auditives (ASA).

On peut cependant se demander si les résultats de Driver ne pourraient pas être compatibles avec des mécanismes d'ASA purs, alliant primitives (mécanismes bottom-up) et schémas (mécanismes top-down), faisant intervenir des schémas (ici audiovisuels) permettant d'identifier une source, et, ainsi, de la filtrer pour mieux identifier la seconde. Ken Grant suit une autre voie, empruntant d'autres paradigmes. Il s'agit, simplement, de tester les capacités de *détecter* le son d'une voix dans un bruit (cocktail party) grâce à la concordance des mouvements visuels. Dans l'expérience princeps [GRA00], on compare une condition de détection sans image (bruit vs. bruit + voix, dans un paradigme de choix forcé à deux intervalles), à une condition de détection avec image (bruit + flux visuel vs. bruit + voix + flux visuel). Les données montrent un gain d'audibilité de la voix en présence des gestes faciaux correspondants (le seuil d'audibilité diminue).

Les données ont été par la suite plusieurs fois répliques, avec une mise en évidence du rôle des corrélations temporelles entre lèvres et mouvements dans la zone F2-F3, et une démonstration de la résistance de cet effet à plusieurs types de dégradations, tant qu'elles ne dénaturaient pas la cohérence naturelle son-lèvres. Ainsi, l'effet résiste à des filtrages passe-bande du son ou au passage en langue étrangère inconnue du locuteur. Mais il résiste mal au remplacement de l'image par des stimuli perturbant l'ordonnancement naturel de la cohérence audio-visuelle : stimuli à l'envers, pour lesquels l'avance naturelle du son sur l'image est contredite ; lèvres remplacées par des ellipses de Lissajous, animées par l'amplitude du son, en perdant là encore l'avance de l'image ; dans ce dernier cas, le gain d'audibilité diminue, sans disparaître.

Il reste que ces expériences, certes claires et convergentes, portent sur la détection et non sur *l'identification* des signaux, ce qui ne permet pas de conclure réellement à l'existence de mécanismes d'extraction précoces avant fusion et prise de décision. C'est là l'enjeu de l'expérience à suivre.

Voir pour mieux entendre, pour mieux comprendre

Dans une série d'expériences que nous avons menées depuis 2 ans avec Frédéric Berthommier et Christophe Savariaux, nous avons cherché une mise en évidence de mécanismes d'interaction précoce en perception de parole. Nous voulions démontrer que, si l'on comprend mieux dans le bruit grâce à l'image, c'est non seulement grâce à un gain *direct* fourni par la lecture labiale, mais aussi grâce à un gain *additionnel* dû à une meilleure audibilité. Nous avons donc décidé de tenter de *geler la voie directe*. L'idée était simple : il s'agissait d'étudier l'identification de stimuli *visuellement identiques* (visèmes) noyés dans du bruit. Nous avons enregistré un locuteur français prononçant des séquences partant d'une position au repos, et allant vers une cible arrondie, typique d'un [y] ou d'un [u], en passant éventuellement par une plosive, dentale ou vélaire, voisée ou non voisée, qui ne perturbe pas le geste labial, soit : [y u ty tu ky ku dy du gy gu]. Essayez de prononcer ces séquences dans un miroir : elles sont visuellement indistingables. Pourtant, nous avons pu montrer que l'intelligibilité audiovisuelle dans un fort bruit était meilleure que l'intelligibilité auditive dans les mêmes conditions de bruit ([SCH04], Exp. 1). Plus précisément, un trait sortait renforcé de l'ajout de la vision : le voisement.

Notre interprétation est la suivante. L'initiation du geste labial commence environ 200 ms avant l'atteinte de la cible, donc le mouvement des lèvres précède d'autant le début de la voyelle. Or le voisement se traduit en français par un prévoisement d'une centaine de millisecondes. Ainsi, la perception visuelle du début du mouvement des lèvres annonce à l'auditeur qu'il doit « tendre l'oreille » pour détecter une éventuelle barre de voisement, si elle existe (Fig. 2). Ce marqueur temporel est susceptible de diminuer le seuil de détection de quelques dBs. Dans une expérience suivante, nous avons confirmé ce résultat, en éliminant toute information résiduelle sur le visage autre que l'information de cohérence lèvres-son (en montant systématiquement sur chaque séquence audio la même séquence visuelle d'arrondissement des lèvres). Le gain d'intelligibilité du voisement était conservé ([SCH04], Exp. 2). Enfin, le remplacement des lèvres par un stimulus visuel non phonétique (une « barre de volume », rouge sur fond noir, croissant et décroissant en parfaite synchronie avec le mouvement des lèvres) a supprimé l'effet ([SCH04], Exp. 3) : l'interaction audiovisuelle précoce est bien démontrée, mais elle semble jusqu'à un certain point « speech specific ».

De l'analyse de scènes à la séparation de sources

Nous avons, à vrai dire, longtemps cherché la trace de mécanismes d'interaction très précoces, et longtemps échafaudé sur les concepts d'Analyse de Scènes de parole Audio-Visuelle (ASAV : voir [BAR98]), avant de parvenir au montage expérimental présenté ci-dessus. Aussi avons-nous eu, les premiers peut-être, l'intuition qu'il pouvait y avoir un corrélat algorithmique à ces interactions précoces, comme la reconnaissance automatique de la parole audiovisuelle est un corrélat algorithmique à la lecture labiale et aux interactions, précoces ou tardives, qui lient le son et l'image dans les mécanismes de compréhension. Ceci a conduit à toute une série de travaux pionniers, d'abord sur le débruitage du son par l'image, dans une approche de filtrage (de Wiener ou LPC), dans les travaux de Girin et al. ([GIR99, 02]; prolongés par Makis Potamianos dans l'équipe de IBM à Watson) .

Puis nous avons abordé le domaine de ce que nous avons appelé la « Séparation de Sources de parole Audio-Visuelle » (SSAV). Dans une configuration classique de séparation de sources, on considère un système possédant N sources inconnues $\boldsymbol{s}$ fournissant, à travers une matrice de mélange elle-même inconnue, P sorties de capteurs $\boldsymbol{x}$. L'opération consiste à estimer la matrice de séparation $\boldsymbol{B}$ dont l'opération $\boldsymbol{y} = \boldsymbol{Bx}$ fournit des estimations $\boldsymbol{y}$ de $\boldsymbol{s}$. Dans les méthodes classiques de séparation aveugle, la matrice $\boldsymbol{B}$ est estimée par un critère de maximisation d'indépendance entre les sorties $\boldsymbol{y}$. Ici nous utilisons une information supplémentaire $\boldsymbol{V_1}$ correspondant aux dimensions géométriques des lèvres d'un locuteur fournies par la vidéo synchronisée avec le signal acoustique $\boldsymbol{s_1}$ que nous cherchons à extraire d'un mélange de sources. Nous avons développé un algorithme utilisant un critère de maximisation de la probabilité conjointe d'observer un spectre et une image du visage associé, et nous obtenons d'excellents résultats de séparation [SOD02, 04]. Parallèlement, Frédéric Berthommier a développé dans notre laboratoire une série d'algorithmes originaux et efficaces, associant son et image (sur la base de corrélations acoustico-visuelles, voir [BAR99]) dans des paradigmes de synthèse visuelle à partir du son, ou de débruitage et resynthèse du son à partir de l'image ([BER03, 04]).

PERSPECTIVE

Nous avons vu, dans notre plaidoyer, pourquoi la multisensorialité était au cœur du processus de la communication parlée, *pour l'usager*, c'est-à-dire pour celui qui parle ou qui écoute. Nous allons maintenant évoquer en quoi elle est au cœur du processus, *pour le chercheur*, en ce qu'elle offre une prise particulière à certaines questions majeures que l'on peut se poser. Les questions récurrentes qui traversent la littérature depuis 50 ans –depuis le départ, pourrait-on dire– sont celles du contenu des représentations perceptives et motrices, de la nature de l'invariance, de la relation entre signaux et code. Nous n'allons pas prétendre répondre ici à ces questions, ni même en présenter une revue exhaustive, mais simplement indiquer comment les propositions les plus récentes intègrent la multisensorialité, et même en quoi la question de la multisensorialité a pu renouveler la perspective.

Formes et mouvements

En ce qui concerne le format des représentations, le débat entre forme et mouvement, posé pour la perception auditive dans les années 70 autour des propositions de spécifications dynamiques de Winifred Strange s'est considérablement renouvelé et a beaucoup avancé dans le domaine de la perception visuelle, sans doute parce que les paradigmes étaient par certains aspects plus simples, mais aussi parce qu'on a pu faire appel à la riche littérature sur la perception visuelle en général. Dans ce secteur, je veux mentionner trois éléments qui me semblent bien baliser le domaine. D'abord, les travaux qui, avec Rosemblum & Saldaña particulièrement, ont repris le paradigme des « Point Light Display » inventé par Johansson et montré ainsi que le mouvement pouvait permettre de spécifier le geste et l'intention, de manière efficace et relativement complète. Mais, en parallèle, il faut mentionner avec force les recherches au long cours de Marie-Agnès Cathiard, qui a accumulé, depuis 15 ans toute une série d'arguments expérimentaux démontrant abondamment –contre ma propre intuition, d'ailleurs !– que le mouvement ne donnait pas toute la clé de l'histoire. Il ne sert que si la forme est insuffisamment spécifiée par le stimulus, ce qui est le cas dans la vision *de face* du geste d'arrondissement, pour laquelle l'ajout d'une composante dynamique permet d'anticiper l'identification du geste ; mais pas en vision de profil, plus efficace, et qui ne bénéficie pas de l'apport d'information visuelle [CAT96]. Il n'est pas toujours donné à entendre, puisque les travaux sur l'anticipation révèlent que le son n'est lâché, dans un geste [i#y], que lorsque les lèvres ont atteint leur cible [CAT96]. Et sa perception « en soi », indispensable pour percevoir visuellement les glides, est délicate [CAT98]. Ceci a conduit Cathiard à proposer une lecture des données en une théorie « shape-from-shading-from-motion » [ABR04] qui s'inscrit bien dans les travaux récents en modélisation neurophysiologique [GIE03].

Perception et action

La question du format des représentations renvoie aussi aux théories de l'invariance, et des relations sensori-motrices. Dans ce contexte, nous avons montré [SCH98] que les données expérimentales semblent plutôt compatibles avec l'architecture de fusion dite RM (Fig. 1d). Or les travaux les plus récents de neurophysiologie semblent confirmer ce schéma, et convergent progressivement vers la mise en évidence d'un circuit cortical de compréhension des actions de l'autre, circuit associant une région temporale de description multisensorielle des caractéristiques de l'action (STS, Sulcus Temporal Supérieur, où convergent des projections des voies auditives et visuelles), une région pariétale postérieure codant la spécification motrice de l'action perçue, et une région frontale inférieure codant/interprétant les buts de cette action, donc les intentions du partenaire [IAC04]. Ce circuit, dans lequel s'inscrivent pour l'essentiel toutes les données d'imagerie cérébrale sur la perception visuelle et audiovisuelle de la parole, apparaît bien comme le support par excellence de l'architecture de fusion la plus compatible avec les données expérimentales.

Le cadre robotique

Reste à imaginer quel pourrait être le cadre de modélisation *global* adéquat. Je prends ici le mot « global » au sens d'une capacité à traiter à la fois des mécanismes de perception et d'action, *sans les séparer mais sans les confondre*. Ainsi, si l'on peut rencontrer dans la littérature des théories perceptives sans action, et même … sans perception, le circuit temporo-pariéto-frontal décrit ci-dessus est à la fois un circuit de l'action, et de la perception des actions, qui laisse la place libre pour l'une comme pour l'autre, dans leur existence propre, et dans leur interdépendance. Dans ce contexte, les recherches que nous avons entreprises avec Jihène Serkhane me semblent très prometteuses. Il s'agit de programmer un *« androïde bébé »,* agent sensori-moteur doté d'actionneurs de la parole (un conduit vocal), de capteurs auditifs, visuels et orosensoriels et de mécanismes de contrôle et d'apprentissage, puis à le mettre en situation d'apprendre à parler, c'est-à-dire d'acquérir contrôles et représentations phonologiques à partir de signaux externes. Cet

agent nous fournit un support théorique et un cadre de modélisation servant de système d'analyse quantitative des données de l'acquisition de la phonétique et de la phonologie. Nous avons pu déterminer comment cet agent doit explorer son espace sensori-moteur entre 4 et 7 mois pour produire des vocalisations compatibles avec celles du bébé au même âge [SER02], puis comment, capitalisant sur les relations sensori-motrices ainsi apprises, il présente des capacités d'imitation précoce similaires à celles démontrées par Patricia Kuhl [SER03]. Or, il n'est pas sans intérêt de rappeler que ces expériences d'imitation sont toujours audiovisuelles : les bébés de 4 mois n'imitent une vocalisation adulte que si elle est présentée dans les deux modalités à la fois [LEB00]. C'est cet ensemble de liens entre audition, vision et motricité, acquis par le bébé robot de nos expériences, qui devrait nous permettre en retour d'examiner comment peuvent se structurer conjointement ces trois modalités, dans le cadre unificateur que nous recherchons depuis le départ.

BIBLIOGRAPHIE

[ABR04] Abry, C. et al. (2004). Some insights in bimodal perception given for free by the natural time course of speech production. In *Festschrift Christian Benoît* (G. Bailly et al. Eds.) : MIT Press (in press).

[BAR98] Barker, J. et al. (1998). Is primitive coherence an aid to segment the scene? *AVSP'98*, 103-108.

[BAR99] Barker, J., & Berthommier, F. (1999). Evidence of correlation between acoustic and visual features of speech. *ICPhS'99*, 199-202.

[BEA96] Beautemps, D. et al. (1999). Hyper-articulated speech: Auditory and visual intelligibility. *Eurospech'99*, 109-112.

[BER03] Berthommier, F. (2003) Audiovisual speech enhancement based on the association between speech envelope and video feature. *Eurospeech'03*, 1045-1048.

[BER04] Berthommier, F. (2004). A phonetically neutral model of the low-level audiovisual interaction. *Speech Communication*, soumis.

[BLO96] Bloch, I. (1996). Information combination operators for data fusion: a comparative review with classification. *IEEE Trans. Systems, Man and Cybernetics, 26*, 52-67.

[BRU03] Brungart, D.S. et al. (2003). Cross-modal informational masking due to mismatched audio cues in a speechreading task. *Eurospeech'03*, 1041-1044.

[CAT94] Cathiard, M.A. (1994). *La perception visuelle de l'anticipation des gestes vocaliques*. Doctorat UPMF, Grenoble.

[CAT96] Cathiard, M.-A. et al. (1996). Does movement on the lips mean movement in the mind? In D. Stork & M. Hennecke (Eds.), *Speechreading by Humans and Machines,* NATO ASI Series pp. 211-219, Springer-Verlag, Berlin.

[CAT98] Cathiard, M.-A. et al. (1998). Visual perception of glides *versus* vowels: The effect of dynamic expectancy. *AVSP'98* 115-120.

[DRI 96] Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature, 381*, 66-68.

[GIE03] Giese, M.A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature*, 4, 179-192.

Le mot de la fin (ou : l'œil et la main)

Ainsi, la multisensorialité est là, ni sujet à part, ni même sans doute sujet en soi, mais partie essentielle du projet, au cœur du processus de la communication parlée : produire des gestes audibles et visibles, pour être entendu et vu ; et comprendre les gestes de l'autre, dans leur multisensorialité naturelle, à travers ses propres capacités d'action. La parole est une boucle, un lien indissociable entre une oreille et une bouche, mais aussi entre un visage et un œil … sans oublier la main … mais ceci est une autre histoire (dont on trouvera quelques prémisses dans les actes du beau colloque « Vocalise to Localise » organisé par Christian Abry et Anne Vilain à Grenoble en janvier 2003 !).

REMERCIEMENTS

A tous ceux, nombreux, qui travaillent ou ont travaillé sur la parole multisensorielle à l'ICP, et dont j'ai pillé sans vergogne les travaux pour les citer ici le plus possible.

[GIR96] Girin, L. et al. (1996). Débruitage de parole par un filtrage utilisant l'image du locuteur : une étude de faisabilité. *Traitement du Signal*, 13, 319-334.

[GIR01] Girin, L., et al. (2001). Audiovisual enhancement of speech in noise. *J. Acoust. Soc. Am., 109,* 3007-3020.

[GRA00] Grant, K.W., & Seitz, P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *JASA*, 108, 1197-1208.

[IAC04] Iacoboni, M. (2004). Understanding others : Imitation, Lnguage, Empathy. In: *Perspectives on imitation: from cognitive neuroscience to social science*, Hurley, S. & Chater, N. (Eds), Cambridge, MA: MIT Press, in press

[JAY] Jaynes E.T. ; Probability theory - The logic of science. Cambridge University Press (in press). http://bayes.wustl.edu (1995).

[LEB00] Lebib, R., & Baudonnière, P.M. (2000). Vocal imitation in 3-month old infants. *Current Psych. Letters*, 2, 79-93.

[MAS98] Massaro, D.W. (1998). *Perceiving Talking Faces*. Cambridge: MIT Press.

[ROB95] Robert-Ribes, J. et al. (1995). A comparison of models for fusion of the auditory and visual sensors in speech perception. *Artificial Intelligence Review*, *9*, 323-346.

[SCH98] Schwartz, J.L et al. (1998). Ten years after Summerfield … a taxonomy of models for audiovisual fusion in speech perception. In R. Campbell, et al. (eds.) *Hearing by eye, II* (pp. 85-108). Hove (UK) : Psychology Press.

[SCH02] Schwartz, J.L. et al. (2002). La parole multimodale : deux ou trois sens valent mieux qu'un. In J. Mariani (Ed.) *Traitement automatique du langage parlé –: reconnaissance de la parole* (141-178). Paris Hermes.

[SCH03] Schwartz, J.L. (2003). Why the FLMP should not be applied to McGurk data. *AVSP03*, 77-82.

[SCH04] Schwartz, J.L. et al. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition*, in press.

[SER02] Serkhane, J. et al. (2002). Etude comparative de vocalisations de bébés humains et de bébés robots**.** *Jeps'2002*, 149-152.

[SER03] Serkhane, J.E. & Schwartz, J.L. (2003). Simulating vocal imitation in infants, using a growth articulatory model and speech robotics. *ICPhS*'2003, 2241-2245.

[SEK 93] Sekiyama, K. & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *J. Phon.*, *21*, 427-444.

[SOD02] Sodoyer, D. et al. (2002). Separation of AV speech sources: A new approach exploiting the AV coherence of speech stimuli. *Eurasip Journal on Applied Signal Processing*, 11, 1165-1173.

[SOD04] Sodoyer, D. et al. (2004). Further experiments on audio-visual speech source separation. *Speech Communication*, soumis.

[SUM 87] Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell (Eds.), *Hearing by eye* (pp. 3-51). Lawrence Erlbaum Associates, London.

[TEI99] Teissier, P. et al. (1999). Comparing models for audiovisual fusion in a noisy-vowel recognition task. *IEEE Trans. Speech and Audio Processing* 7, 629-642.

[TII01] Tiippana, K. et al. (2001). Visual attention influences audiovisual speech perception. *AVSP'2001,* 167-171.

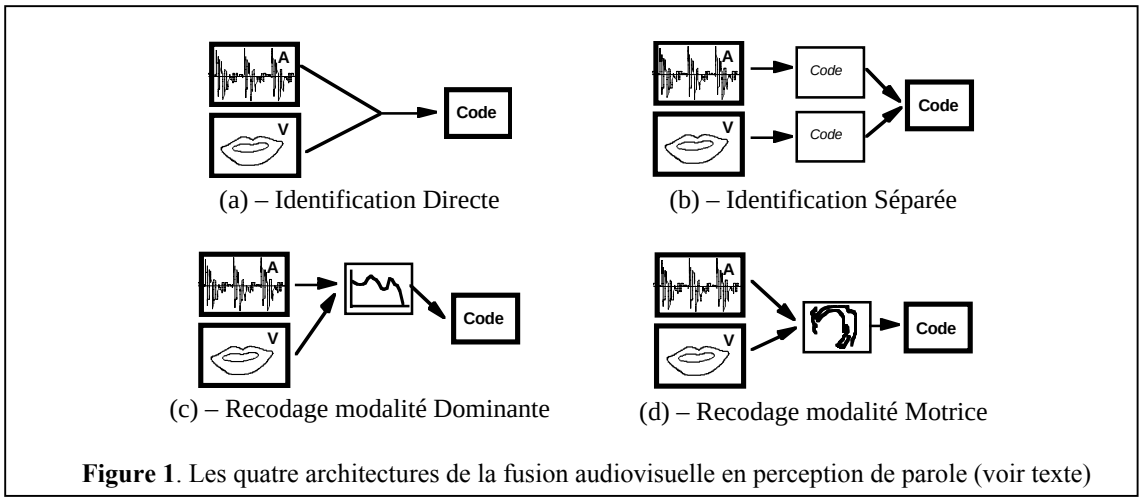

Figure 1. Les quatre architectures de la fusion audiovisuelle en perception de parole (voir texte)

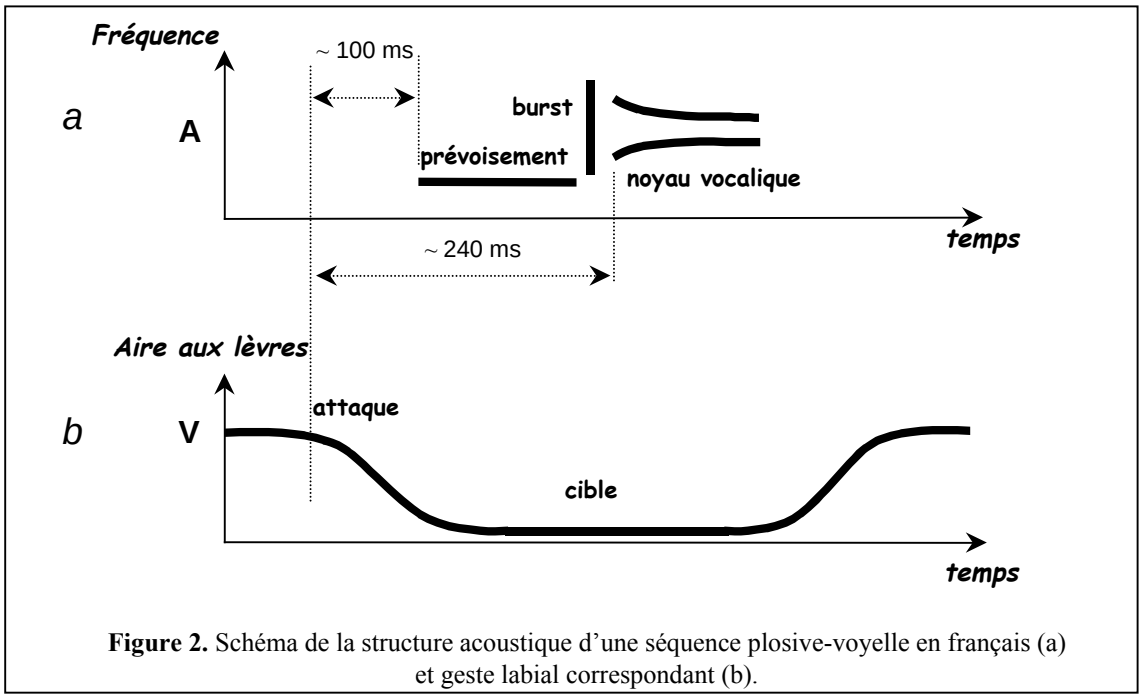

Figure 2. Schéma de la structure acoustique d'une séquence plosive-voyelle en français (a) et geste labial correspondant (b).