Modèles évolutifs pour la perception artificielle des sons et des images

Harouna KABRE, Jean CAELEN & Anne SPALANZANI

Laboratoire CLIPS/IMAG,
Domaine universitaire
BP 53, Bat B, 38041, Grenoble Cedex 9
Harouna.kabre@imag.fr
Jean.Caelen@imag.fr

Résumé

Cet article décrit un modèle unifié pour l'intégration d'un niveau perceptif dans les systèmes d'interaction homme-machine multimodaux. Cette recherche a pour objectif de rendre ces systèmes plus robustes. Nous suggérons que la robustesse résulte de la combinaison d'un apprentissage local (au moyen de réseaux connexionnistes) et d'un apprentissage global (fondé sur des algorithmes génétiques). Nous aboutissons ainsi une meilleure prise en compte par le système des perturbations liées à un environnement audiovisuel changeant ; cela conduit également à une adaptation en continu du système à son environnement. Le système est démontré à l'aide de quelques résultats.

Mots clefs : Perception artificielle, Interaction homme-machine, Multimodalité.

1. Introduction

On sait que des sources d'information extra-linguistiques - comme la gestuelle du locuteur, ses mimiques ou expressions faciales - participent au processus de communication verbale en améliorant la compréhension de l'allocutaire. La perception du mouvement des lèvres du locuteur, mouvement synchrone au son produit, participe également à la robustesse de la compréhension de l'allocutaire dans un environnement bruité (effet cocktail party). Mais on sait aussi que ces mêmes sources d'information peuvent perturber la compréhension si elles ne sont pas synchrones (par exemple le retard introduit entre la voie vidéo et la voie audio sur une bande cinématographique) ou si elles sont contradictoires. Il est donc clair que ces informations influent dans un sens ou dans un autre sur la compréhension globale des énoncés verbaux (effet MacGurk). Nous faisons l'hypothèse, qu'une machine équipée de capteurs (caméra par exemple) peut mettre à profit ces sources d'information pour améliorer la robustesse de l'interaction entre un usager et une machine à condition de bien percevoir les signaux qui concourent au sens du message. Perception et sens sont évidemment liés.

Dans la chaîne des traitements qui conduisent à la compréhension d'un discours oral (ou d'une conversation), le niveau perceptif est semble-t-il le premier niveau à faire émerger des représentations favorisant l'ancrage du sens. Or le niveau de représentation audiovisuel a été peu étudié en regard des situations de communication et du langage : c'est ce niveau de catégorisation pré-symbolique précoce (avant la catégorisation lexicale et peut-être même phonétique), qu'il semble intéressant d'étudier pour faire avancer les connaissances sur la communication multimodale. Il est évident qu'il y a lieu d'étudier également la voie descendante qui, des niveaux

symboliques (sémantique surtout), apporte des informations qui guident la perception. Combien de fois a-t-on cru entendre quelqu'un dire cela, alors que l'on ne pouvait pas l'entendre du fait du bruit excessif ? Ainsi l'influence des niveaux les uns sur les autres donnent-ils lieu à de nombreux phénomènes (anticipation, reconnaissance biaisée, ambiguïtés, conflits, effets de masque, etc.) qu'il y a lieu d'étudier pour mieux comprendre où et quand émerge le sens, comment et par quoi le système de compréhension est fragilisé ou solidifié.

Formulé quelque peu différemment, le but de notre article est donc de contribuer à l'approfondissement des connaissances sur le rôle du niveau perceptif (multisensoriel) en interaction multimodale (dans un premier temps en reconnaissance audio-visuelle de la parole), en expérimentant et en simulant un tel niveau de manière logicielle. La méthode générale sera donc de partir de données verbales enregistrées en situation de dialogue humain (ou homme-machine), selon des scénarios fortement contrôlés, de traiter ces données avec des méthodes objectives (classification automatique, catégorisation par réseaux de neurones, reconnaissance automatique, etc.) et d'en tirer des enseignements en regard des théories actuelles, en particulier en bénéficiant des derniers apports de la *vie artificielle*. Evidemment, ce travail ne peut pas apporter de connaissances directes sur les processus humains mis en jeu, il vise simplement à montrer à l'aide d'artefacts, l'apport d'un niveau « perceptif » clairement identifié dans un système d'interaction homme-machine multimodal.

2. Point de vue des sciences de la cognition

Plusieurs approches théoriques traitant des rapports de la perception et du sens ont été proposées dans le domaine des sciences cognitives.

a) Les cognitivistes (Fodor, 1983; Marr, 1982) ont en commun une théorie centrale : la computation reste indépendante des objets manipulés. Par contre ils diffèrent entre eux sur l'idée qu'ils ont des représentations : fragment d'information stocké, processus de structuration de la connaissance, propositions logiques, états mentaux ou réalisation physique (Marr, 1982), etc., ainsi que sur les théories calculatoires sous-tendues par ces représentations. Ils placent souvent la syntaxe avant la sémantique dans la mesure où la syntaxe d'un langage formel qui règle la bonne formation des symboles, est une computation sur ces symboles, et où, dans un deuxième niveau, ces symboles sont mis en correspondance avec les objets du monde par une relation sémantico-pragmatique. Raisonnant ainsi de niveau à niveau, ils proposent une approche descendante jusqu'au niveau perceptif traité de processus "irrépressible" par Fodor.

b) Le constructivisme possède des racines historiques anciennes (les sceptiques grecs, Whitehead, la Gestalt, Paillet, etc.) et constitue un courant qui n'a pas de cohérence intrinsèque. L'idée fondamentale est de considérer la cognition dans sa genèse, comme phénomène naturel, en tant que les espèces se déterminent par rapport à leur environnement. La cognition est donc « savoir-faire » (capacité notamment de savoir se maintenir en vie) dans lequel la connaissance est à la fois un processus opératoire et l'expression de ce processus. Dans cette perspective, la perception est première et le sens en est un phénomène dérivé.

c) Plus récemment, le néo-constructivisme fait de l'organisation de la vie elle-même, la base de la cognition (Varela, 1989). Le paradigme essentiel de cette approche est que c'est la cohérence interne de l'organisme qui en caractérise l'identité et dirige l'adaptation. Cette dernière est donc invariable. Les systèmes se créent, évoluent, s'adaptent, se modifient. L'outil de simulation privilégié par les néo-constructivistes est le réseau neuro-mimétique, bien qu'il soit incapable de reproduire les fonctions du vivant.

Ces trois grands courants scientifiques que l'on peut ramener aux deux courants cognitiviste et constructiviste de part leur divergence de conception du monde - conception guidée par les stimulus pour constructivistes et conception guidée par les connaissances - s'affrontent sur le terrain de l'apprentissage. Il s'agit de savoir si la notion d'apprentissage doit ou non être basée sur un conditionnement de réflexes par manipulation externe de liens entre stimulus et réponses (Pavlov) ou sur la genèse des formes représentationnelles d'orientation symbolique (Vigotski, 1934) ; en terme informatique si cette notion d'apprentissage repose sur des modèles numériques ou des modèles symboliques ou encore sur l'apprentissage supervisé ou non supervisé. Des tentatives récentes, loin d'apaiser la controverse sur le plan théorique, proposent une nouvelle interprétation "intégrationniste" des deux tendances sur le plan de l'implémentation computationnelle (Forest & Siksou, 1994).

Outre les problèmes théoriques posés sur la nature des mécanismes d'apprentissage, **les conditions et le contexte** dans lesquels l'apprentissage se développe sont presque absents dans la plupart des travaux actuels. Pourtant plusieurs études notamment dans le domaine de la perception audiovisuelle ont permis de déterminer des conditions environnementales qui peuvent contribuer plus fortement à la prise de décision en perception audiovisuelle. D'après Summerfield (1983) les images qui donnent le plus d'informations sont celles :

- qui se situent à une distance de 1.5 m et qui sont bien éclairées,
- qui montrent le corps et les bras du locuteur,
- où le visage n'a pas de moustache ni de barbe,
- et qui montrent les lèvres maquillées pour mettre en valeur la forme des lèvres.

Pour qu'une image puisse jouer un rôle dans la perception audiovisuelle, il faut qu'elle soit écologiquement pertinente. Par exemple, si on présente un son avec l'image du locuteur à l'envers (le haut vers le bas), l'image n'affecte pas ou peu la perception de la parole. Mais si on retarde quelque peu les mouvements des lèvres par rapport au son sur une bande cinématographique, alors les conséquences peuvent être fâcheuses sur la compréhension de la parole.

Posé dans un champ plus général, notre problème concerne la prise en compte de différentes sources d'informations audiovisuelles et, par simulation, d'en mesurer les effets sur la compréhension d'un être humain qui percevrait des signaux (visuels ou acoustiques) plus ou moins masqués les uns par les autres. Cet être humain est capable d'en extraire les caractéristiques essentielles même si les conditions environnementales deviennent défavorables (environnement bruyant ou mal éclairé) et d'apprendre à généraliser ses savoir-faire en reconnaissance des signaux. C'est cet aspect de l'apprentissage que nous tentons de modéliser dans le domaine de la reconnaissance audiovisuelle de la parole avec l'hypothèse que de tels modèles devraient contribuer à alimenter le débat fondamental sur l'apprentissage et à nous apporter des nouvelles solutions pour l'étude de la robustesse des systèmes artificiels.

3. Modèle PDM

Pour introduire la notion de contexte dans le processus d'apprentissage et d'adaptation à l'environnement à la fois chez l'émetteur et chez le destinataire, nous considérons que la communication entre humains mettrait en œuvre, chez l'émetteur, un processus de sélection d'opérateurs qui au final produirait des signaux sonores et visuels utiles au processus d'échange d'information. On fait l'hypothèse que les signaux émis tiennent compte du contexte et des conditions d'émission que l'émetteur aurait pris en charge au niveau cognitif et au niveau perceptif à l'adresse du destinataire (principe de pertinence [Sperber et Wilson, 86]). Les sons et les images

produits seraient ainsi le résultat de cette sélection d'opérateurs tant au niveau de la production que du sens à véhiculer ; puis ils sont soumis à des perturbations de sources diverses (bruit acoustique de l'environnement pour les sons, variation d'éclairage pour les images, etc.). La tâche du destinataire, placé à une certaine distance du locuteur, qui cherche à analyser ces signaux, est donc quelque peu différente : il s'agit pour lui (a) de se mettre dans les conditions perceptives optimales et (b) de trouver les opérateurs sélectionnés par l'émetteur pour décoder le sens ou de sélectionner ses propres opérateurs en fonction de ses attentes. Dans PDM (Perceptive to Dialogue Model), nous tentons de modéliser le destinataire dans sa fonction de reconnaissance de messages audiovisuels, comme un processus inverse de celui de l'émetteur : nous supposons qu'il tente de sélectionner au mieux des opérateurs qui lui permettent de décoder le signal reçu. Ce problème est équivalent à considérer une population « d'organismes » effectuant chacun une opération puis à sélectionner l'individu le plus performant. On peut ainsi se ramener à un problème d'optimisation utilisant les principes de la vie artificielle. Chaque « organisme » est un réseau de neurones (RN) soumis à un processus de sélection par un algorithme génétique.

La modélisation des perturbations dans PDM consiste à générer diverses perturbations (bruit acoustique, variation de lumière, etc.) qui convoluées avec les signaux sonores et visuels préalablement enregistrés sur un locuteur permettent de modéliser différents contextes d'apprentissage. Nous disposons donc d'une formulation générale du problème, du simulateur EVERA (voir paragraphe suivant) qui permet de plonger des « organismes » dans un environnement audio-visuel et de sélectionner les plus robustes. C'est cette procédure de sélection qui permet (a) de remonter à la sélection des opérateurs décrits ci-dessus et (b) de garder les organismes les mieux placés au sens d'un critère perceptif. Les paragraphes suivants détaillent ces points.

4. Cadre expérimental

Dans notre simulateur EVERA (Environnement de Vie artificielle pour l'Etude de la Robustesse des Apprentis), décrit plus en détail dans (Kabré & Spalanzani, 1996), l'environnement simulé est un modèle mathématique d'une pièce (4 murs, un plafond et un sol) qui permet de faire varier les dimensions de cette pièce et ses coefficients de réverbération acoustique ; il est possible de choisir également les sources de perturbation (les différents bruits qui peuvent exister dans cet environnement acoustique et le niveau d'éclairage). Les "organismes" plongés dans cet environnement représentent une population de systèmes de reconnaissance automatique de la parole (chaque système utilise une technique connexionniste) qui évoluent et se déplacent dans cet environnement. Ils ont des capacités de perception, d'apprentissage, et de reproduction, capacités mises en œuvre en trois étapes: extraction de paramètres des signaux, intégration des paramètres et adaptation des organismes. Les organismes se distinguent entre eux par une paramétrisation différente de leurs réseaux connexionnistes. Ils ont donc des performances différentes pour une tâche identique de perception puis de reconnaissance. Ce sont sur ces différences que reposent la sélection et l'amélioration de la population.

Chaque organisme se présente donc selon le schéma de la fig. 1, il reçoit une information multimodale (parole et visage parlant), la traite et la fusionne aux niveaux son et image, l'identifie en s'adaptant le mieux possible aux conditions de l'environnement qui évoluent dans le temps. D'autre part, l'ensemble de la population évolue par mutation génétique pour en optimiser l'adaptation générale.

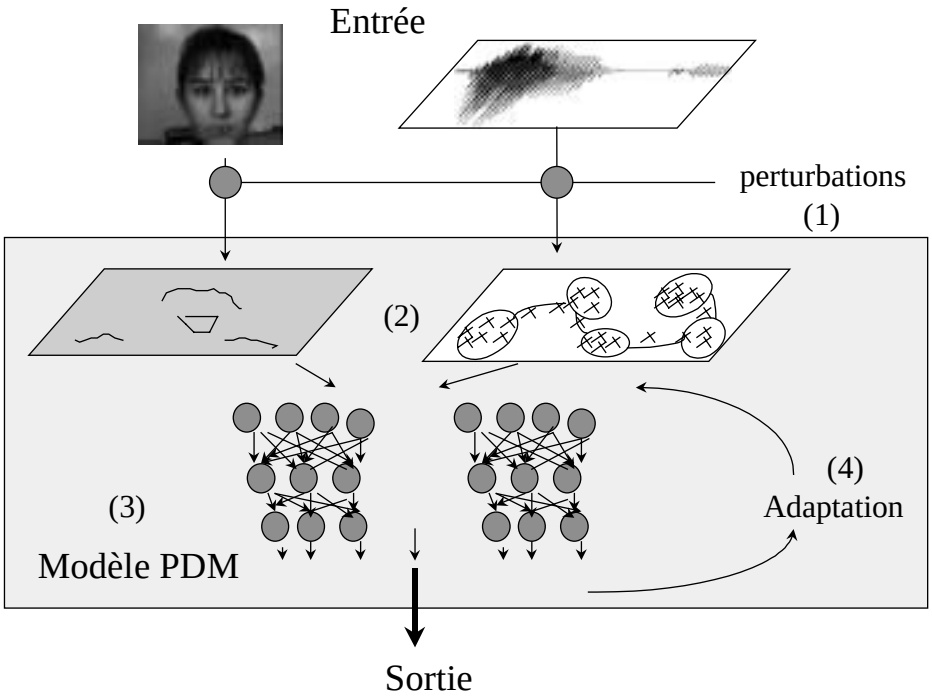

Fig. 1 : Le modèle PDM : extraction d'informations audiovisuelles (1,2) soumises à perturbations, intégration et classification par réseaux de neurones (3) et adaptation aux variations de l'environnement d'apprentissage par sélection (4).

4. 1. Extraction des paramètres audiovisuels

L'environnement considéré est une pièce virtuelle dans laquelle on peut calculer en chaque point de celle-ci les sons propagés selon ses propriétés physiques (taille, volume, coefficient de réverbération, etc.) ainsi que les images d'un visage de différentes tailles (selon l'éloignement). Les signaux résultant d'un mélange son-image, sont modélisés en tous points de la pièce dans laquelle sont placés les organismes. Chaque organisme possède un analyseur acoustique, un analyseur d'images et un réseau connexionniste de classification jouant le rôle de système de reconnaissance. La phase d'analyse a pour fonction d'extraire du signal une information spectrale et une information de contours de lèvres (essentiellement le degré d'ouverture) utilisables en entrée du réseau connexionniste de classification des informations multimodales. Dans cette étude les PLP (Perceptual Linear Predictive coefficients (Hermansky, 1990)) sont utilisés pour la caractérisation des sons et le degré d'ouverture des lèvres (Kroschel, Kabré, 1996) pour les images.

4. 2. Reconnaissance audio-visuelle

La reconnaissance s'effectue par des réseaux à propagation du gradient. Il y a 10 cellules sur la couche d'entrée, 30 sur la couche cachée et 10 en sortie. Les cellules d'entrée reçoivent 7 coefficients PLP et 3 paramètres d'ouverture des lèvres. Les 10 cellules de sortie sont spécialisées pour la reconnaissance des 10 voyelles du français (voir plan d'expérience ci-dessous). L'intégration des données audio-visuelles s'effectue donc par l'intermédiaire de la couche d'entrée puis par l'ensemble du réseau. Lors de la phase d'adaptation du réseau la différence entre la sortie calculée et la sortie désirée permet de calculer une erreur de reconnaissance. Cette dernière permettra de donner un coefficient de *fitness* (degré d'adaptation à l'environnement) nécessaire à l'étape de sélection.

Chaque "organisme" a un réseau d'architecture identique. Les réseaux ne différent entre eux que par les poids de leurs connexions. Ces poids ont donc un effet indirect sur la contribution des sources acoustiques et visuelles et c'est par leur intermédiaire que la sélection perceptive pourra jouer.

4. 3. Adaptation des organismes

Dans notre système, les organismes sont caractérisés par leur génotype qui est représenté par une liste de poids de connexions des neurones de leur réseau. Ce génotype permet d'établir une note de performance de catégorisation. Les organismes ayant les génotypes les plus adaptés, qui auront une faculté plus grande de catégoriser les entrées, seront mieux classés que les moins performants.
Par la méthode des algorithmes génétiques, qui consiste à sélectionner et muter les meilleurs organismes, nous faisons évoluer la population d'organismes vers une meilleure performance globale. La séquence d'instructions pour l'évolution des organismes est la suivante:

0. Acquérir une donnée,
1. sélectionner deux "bons" organismes,
2. effectuer le croisement entre ces deux organismes,
3. effectuer des mutations sur les deux "nouveau nés",
4. remplacer dans la population les deux parents par les deux "enfants",
5. retour à 0.

La sélection s'opère selon un coefficient de *fitness* (i.e. degré d'adaptation à un environnement) qui est calculé en fonction de l'erreur de perception des sons émis. Les organismes ayant le réseau le plus adapté à la reconnaissance des sons à catégoriser auront une probabilité plus grande de se reproduire que celles qui sont moins adaptées et qui ont un taux d'erreur plus grand. Le croisement permet d'interchanger une partie des poids de chacun des organismes de la manière suivante:

1. choix de l'endroit de la césure du génotype des parents (père et mère),
2. croisement de la partie droite du père avec la partie gauche de la mère,
3. croisement de la partie droite de la mère avec la partie gauche du père.

La mutation permet d'altérer quelques poids selon une fonction aléatoire.

Cet ensemble d'instructions appliqué à chaque génération d'organismes permet d'améliorer la population de réseaux de neurones en terme de taux de reconnaissance de signaux acoustiques et visuels.

5. Résultats

Nous avons mené nos expériences sur deux populations d'organismes : l'une (RNGA ou population « évolutive ») apprend en s'adaptant et l'autre (RN ou population « non-évolutive ») s'adapte uniquement aux changements de l'environnement. Deux conditions d'adaptation sont considérées :

(a) pour la première, la population initiale subit un apprentissage préalable sur un ensemble de sons d'entraînement (50 répétitions de 10 voyelles du français émises dans une pièce virtuelle de coefficient de réverbération de 0.6 en présence d'un bruit gaussien à -6 dB). Puis un ensemble de sons-tests (un autre jeu de répétition des 10 voyelles) leur est présenté à 3 autres valeurs de rapport signal à bruit (-6, 0, 6, 12, 20 dB).

(b) pour la deuxième, la population initiale ne subit aucun apprentissage préalable ce qui correspond à une tâche nettement plus difficile. Pour ces populations « évolutives » nous avons fait varier l'ensemble des paramètres habituels des algorithmes génétiques (taux de mutation, taux de croisement et taille de la population). Les populations « évolutives » et « non-

évolutives », nous sont testés successivement dans deux environnements de complexité croissante. Le premier n'intègre que des perturbations limitées (bruit acoustique uniquement). Le deuxième intègre du bruit et de la réverbération. Les bruits arrivent soudainement dans l'environnement, ce sont des sons de porte qui claque, de réveille-matin, de klaxon, etc. Les figures 2 et 3 montrent les performances de ces populations dans ces deux environnements avec et sans pré-apprentissage (en abscisse le temps annoté des changements d'environnement et en ordonnées le taux global de reconnaissance des voyelles).

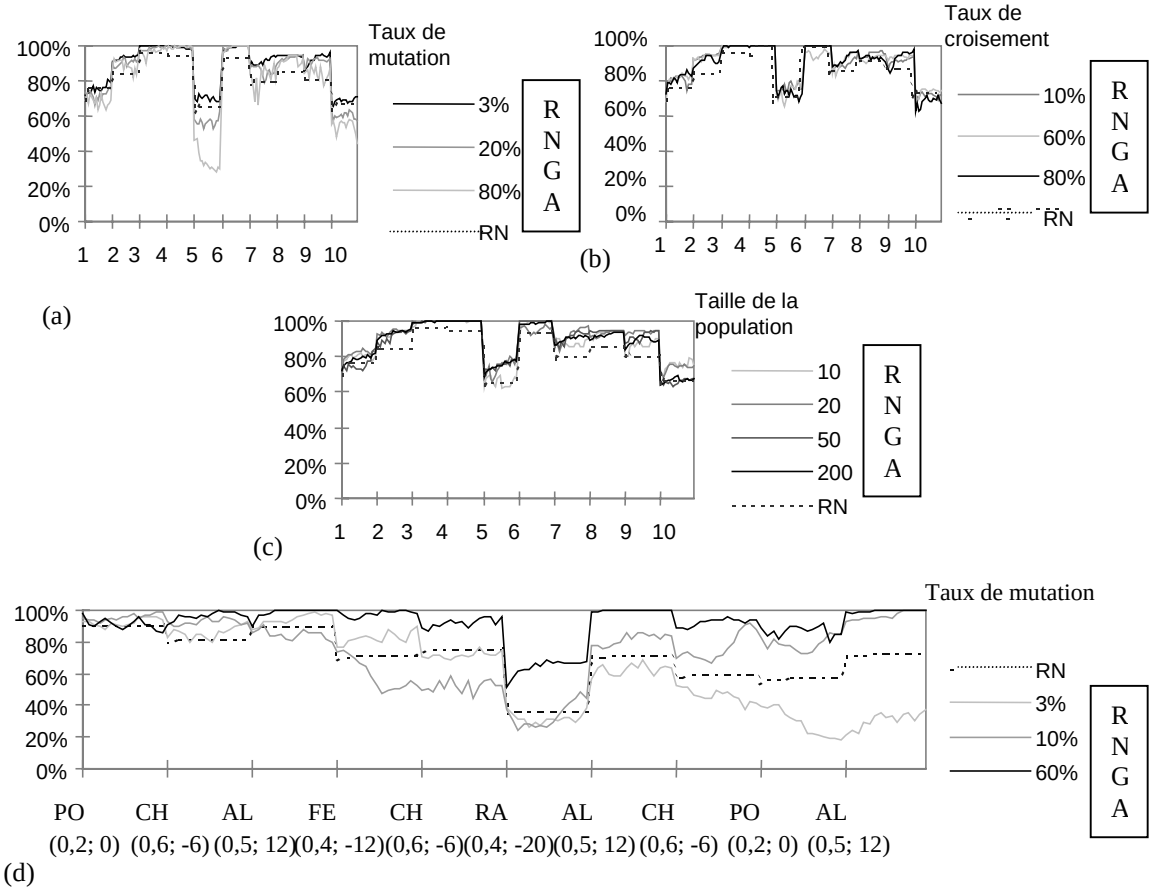

Fig. 2 : Adaptation de différents types d'organismes aux changements de l'environnement acoustique dans le temps. Les organismes n'intégrant pas des capacités d'évolution (RN) sont comparées à celles intégrant pas la composante évolutive (RNGA) pour différentes valeurs des paramètres d'évolution tels le taux de mutation (a), le taux de croisement (b), la taille de la population (c) et pour un environnement complexe (bruit et réverbération) (d). En abscisse le temps gradué selon les numéros des instants de changement d'environnement (la durée de chaque incrément vaut 15 générations) ou avec le nom des changements acoustiques qui ont été opérés. Pour les figures (a), (b), et (c) ces changements correspondent à un environnement acoustique simple dont seul le rapport signal à bruit change ; pour (d) la nature du bruit, le taux de réverbération et le rapport signal à bruit sont notés. Ainsi (PO, 0.2 ; 0) correspond à un bruit de porte avec un taux de réverbération de 0.2 et un RSB de 0 dB. Les autres bruits sont : AL alarme, RA radio, FE bruit de feu, CH bruit de chaise. Pour toutes ces figures en ordonnée figure le taux global de reconnaissance des voyelles. Lors des changements brusques d'environnement acoustique, on note une chute de performance plus faible dans les populations « évolutives » RNGA qui se montrent donc plus résistantes à ces changements.

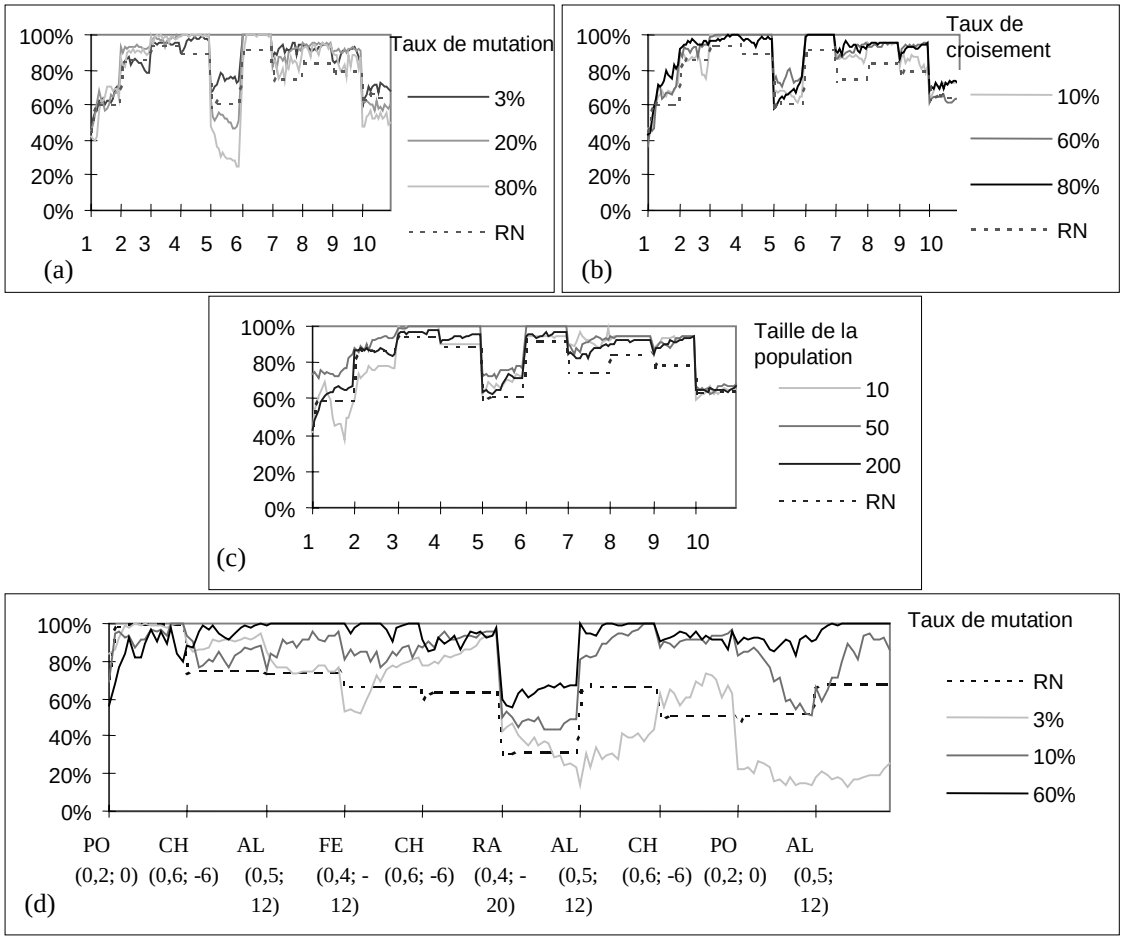

Fig. 3 : Même légende que sur la Fig. 2 mais en commençant l'évolution sans pré-apprentissage de la population initiale. Malgré la dégradation du taux d'adaptation les populations « évolutives » RNGA parviennent à s'adapter. On note une performance moyenne des RN autour de 70% (d) ce qui confirme le fait que ces techniques n'apprennent pas bien en situation de changement rapide dans l'environnement.

On peut noter que dans les deux cas les populations « évolutives » RNGA parviennent à mieux s'adapter aux changements de l'environnement que l'autre population RN. Pour les faibles variations de l'environnement (fig. 2a,b,c) les deux populations obtiennent globalement le même taux d'adaptation (sauf pour certains RNGA à taux de mutation de 80%). Lors des changements brutaux d'environnement acoustique, on note une chute de performance plus faible pour les populations RNGA dont le bon réglage des paramètres se situe à un taux de mutation de 10% pour une population initiale de 50 organismes et un taux de croisement de 50%.

On remarque également que lorsqu'un environnement a déjà été « vu » par un organisme, même dans des conditions légèrement différentes (eut égard au rapport RSB ou à la valeur de réverbération), il s'adapte plus vite et obtient des scores de reconnaissance meilleurs (c'est le cas par exemple sur la fig. 3 pour le bruit AL présenté une deuxième fois au temps 7 pour la population RNGA mutée à 10%). Ce n'est pas toujours vrai cependant pour des populations à fort taux de mutation (dans ce cas la « mémorisation » du contexte se fait moins sentir). Il est évident aussi que la stabilité des performances de reconnaissance au cours de la mutation est plus grande pour une deuxième présentation du même environnement (c'est particulièrement visible sur la fig. 3 entre le début de la session, bruit PO pour la population RNGA mutée à 60% et le même bruit présenté au temps 9 à la même population).

En résumé, la rapidité de l'adaptation est plus grande sur une population à taux important de mutation. Cette population peut en revanche perdre sa capacité de mémorisation des situations déjà vécues car l'effet de rémanence est évidemment moins important. Il s'agit donc pour chaque application de trouver un compromis entre ces deux effets.

Ces résultats montrent en outre, qu'un cadre de vie artificielle permet d'étudier des phénomènes dont la complexité ne semble pas facile à maîtriser suivant d'autres méthodologies.

6. Conclusion

Cet article a proposé un cadre *de traitement perceptif de l'information multimodale* dans lequel une introduction contrôlée de *contraintes* venant à la fois des niveaux supérieurs (sélection des opérateurs) et de l'environnement permet d'améliorer la robustesse des systèmes de reconnaissance audiovisuelle de la parole. On a montré quelques résultats sur l'apprentissage en présence de perturbations (bruit, réverbération, etc.). Dans les expériences menées le modèle d'interaction entre l'apprentissage et l'environnement fondé sur la combinaison d'un apprentissage local (réseaux de neurones) et d'un apprentissage global (algorithmes génétiques) a fourni les meilleurs taux d'adaptation. La robustesse résulte de cette combinaison. Nous aboutissons ainsi une meilleure prise en compte par le système des perturbations liées à un environnement audiovisuel changeant ; cela conduit également à une adaptation en continu du système à son environnement.

Ainsi ces résultats renforcent l'idée que les conditions environnementales et le contexte influent sur la qualité de l'apprentissage. En outre lors des grandes dégradations de l'environnement, les pertes en performances sont moins marquées sur la population qui apprend tout en s'adaptant.

7. Bibliographie

Adler D., L'apprentissage dans les sciences cognitives: approches théoriques, Apprentissage et Machines- Intelletica, 1987.

Fodor, J.A. (1983) The modularity of Mind. An essay on faculty psychology. Cambridge (Mass.), MIT Press.

Fourcin, A., Progress Overview of The SAM Project, proceedings of EUROSPEECH, Paris, 1989.

H. Kabré, Performance and Competence Models for Audio-Visual Data Fusion, SPIE'95, 1995.

H. Kabré, On the Active Perception of Speech By robots, IEEE/RJ Multi-sensor Fusion and Integration for Intelligent Systems, Washington D. C., pp. 765-774, 1996.

K. Kroschel , M. S. Mekheil & H. Kabré Modelling the Lip-Contour in Noisy Images for Lipreading Applications, Institut Franco-Allemand d'Automation, Karlsruhe, 1996.

H. Kabré & A. Spalanzani, EVERA: A system for the Modelling and Simulation of Complex Systems, First Int. Workshop on Frontiers in Evolutionnary algorithms, FEA'97, USA, North Carolina, pp. 184-188, USA, 1997.

Pierrel, J.M., Informations lexicales dans un dialogue homme-machine, Séminaire lexique,IRIT-UPS Toulouse, pp. 1-20, 1992.

Poggio T., Low-level vision as inverse optic, Symposium : Computational model of hearing and vision, Tallinn, pp. 123-127, 1984.

Varela, F.J. (1989) Autonomie et connaissance. Editions du Seuil, Paris.

H. Kabré & J. Caelen, Communication verbale et non-verbale : vers une intégration de la perception artificielle et de l'interaction multimodale, AIDRI97, Apprentissage : des principes naturels aux principes artificiels, 1997.