UNIVERSITÉ DE GRENOBLE

# THÈSE

Pour obtenir le grade de

## DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Sciences du Langage**

Arrêté ministériel : 7 août 2006


Présentée par

## Anne VANPÉ

Thèse dirigée par **Véronique AUBERGÉ**

préparée au sein du **Laboratoire GIPSA-lab**
dans **l'École Doctorale Langues, Littératures et Sciences Humaines**


# Expressions et micro-expressions spontanées de la face et de la voix en Interaction Homme-Machine : esquisse d'un modèle du "Feeling of Thinking".

Thèse soutenue publiquement le **21 avril 2011,**
devant le jury composé de :

**M. Philippe MARTIN**
Professeur à l'Université Paris Diderot 7, Rapporteur
**Mme Catherine PELACHAUD**
DR CNRS au LTCI, Telecom Paris Tec, Rapporteur
**M. Didier DEMOLIN**
Professeur au GIPSA-lab, Université Stendhal, Président
**Mme Véronique AUBERGÉ**
CR CNRS, GIPSA-lab, Université Stendhal, Directrice de thèse
**M. Jean-François BONASTRE**
Professeur au LIA, CERI, Examinateur
**M. Jean-Marc COLLETTA**
Professeur au LIDILEM, Université Stendhal, Examinateur

# RÉSUMÉ / ABSTRACT

Les technologies de l'interaction personne-machine se centrent de plus en plus sur l'humain. Le statut informatif des expressions présentes massivement en dehors des tours de parole, dans les micro-événements visibles ou audibles, inscrit le sujet dans une communication permanente du *Feeling of Thinking* (états physiologiques, mentaux, émotionnels, intentionnels et attitudinaux). Ce travail a été mené sur un corpus induit émotionnellement et restreignant l'interaction personne-machine au minima langagier. Une méthodologie empirique, sur des principes éthologiques, a d'abord été construite pour annoter les micro-gestes audibles et visibles de 6 sujets. Une analyse perceptive a ensuite mesuré la pertinence communicative de certaines icônes gestuelles. Puis a été menée une étude des occurrences des micro-gestes dans l'organisation temporelle de la tâche et des tours de parole, traçant ainsi le comportement des sujets. Enfin a été proposée une qualification impressionniste des nombreux sons vocaux non lexicaux qui ponctuent les performances (bruits de bouche, *grunts*, *bursts*, *fillers*, interjections).

**Mot-clés** : *Feeling of thinking*, micro-gestes expressifs, interaction multimodale, interjection, backchannel, prosodie expressive.

Communicative interaction technologies focus more and more on human aspects. The informative position of expressions that are massively present out of the talk turns, through visible and audible micro-events, place the listener/interlocutor in a continuous communication of "Feeling of Thinking" (physiological, mental, emotional, intentional and attitudinal states). This work has been carried out on an emotionally induced corpus that greatly limits speech of the human-machine interaction. An empirical methodology based on ethological principles has been built to annotate audible and visible micro-gestures by 6 subjects. A perceptive analysis has measured the communicative relevance of certain gestual icons. A study of micro-gesture occurrences in the temporal organisation of the task and within the turn-taking provides a cue of subjects' behaviour. Finally, this study proposes an impressionistic characterisation of the numerous vocal non-lexical sounds disseminated within the performances (mouth noises, grunts, fillers, interjections).

**Keywords** : *Feeling of thinking*, expressive micro-gestures, multimodal interaction, interjection, backchannel, expressive prosody.

# REMERCIEMENTS

Je remercie tout d'abord Véro, ma directrice, sans qui travailler sur ce sujet vaste et passionnant n'aurait pas été possible. Travailler avec elle m'a obligée à acquérir patience, adaptation et gestion du stress, mais surtout l'autonomie, la rencontre des autres et la « débrouille », et je la remercie pour cela.

Didier, que dire... Un grand merci pour ton engouement pour la recherche et l'étendue de tes connaissances. Merci surtout pour ta présence, tes conseils et ton soutien, et pour avoir accepté d'être mon président de jury.

Merci à Catherine Pelachaud et Philippe Martin pour avoir accepté de rapporter cette (longue) thèse. Plus globalement, merci à tous les membres de mon jury d'avoir répondu présent, et de vous être intéressés à mon travail. Je remercie en particulier Catherine Pelachaud, Jean-Marc Colletta, Christian Graff et Nick Campbell, pour les quelques discussions scientifiques que nous avons pu avoir à un moment ou un autre de ma thèse, et qui m'ont aidée et / ou influencée scientifiquement.

Je remercie mes sujets « piégés » par SoundTeacher, en particulier « T » et « S », et les dizaines de juges de mes tests perceptifs, d'avoir accepté de passer du temps à « servir la science » ;-) . Sans eux mes recherches n'auraient pas été possibles.

Merci à mes correcteurs d'un jour ou de toute une thèse (collègues, famille ou amis qui se reconnaîtront) d'avoir relu mes énormes chapitres et mes phrases trop longues, d'avoir corrigé coquilles, fautes de frappe, et « fautes bêtes », et couper mes phrases incompréhensible en 2, 3 ou 4... Mention spéciale à Ninie et maman, qui ont presque lu entièrement mes 300 pages et quelques, sans être spécialistes du domaine.

Merci aux « techniques » du labo, Christophe, Lionel, Pipou et aussi Alain, qui par leur connaissances des différents outils informatiques, acoustiques et linguistiques, et aussi par leurs astuces en tout genre (et hop! Un petit script!!), ont permis la résolution de bien des problèmes, et en bien moins de temps que ce qu'on aurait pu imaginer.

Merci aux administratives : Dominique, Marie-Thé, Houria, et celles qui sont restées moins longtemps, pour leur gentillesse, leurs attentions et leur disponibilité.

Merci à ceux de l'équipe, toujours présents au labo ou dont la vie les a menés ailleurs, qui ont supporté mes humeurs et mon dynamisme au jour le jour : Albert et Nico, qui ont toujours répondu présents pour m'aider et répondre à mes questions techniques, statistiques ou scientifiques, mais aussi pour leur patience, leurs taquineries et pour leur caractère zen et agréable ; la discrète et adorable Fanny qui a déblayé le terrain du

*FoT* et fait une bonne partie du « travail ingrat » d'étiquetage ; et aussi Taka, Kévin, Rosario, Lu Yan la petite dernière, pleine de petites attentions, et tout ceux que j'oublie.

Merci à toutes les personnes du site Stendhal du laboratoire, permanentes ou de passage au cours des quelques 6 années passé au sein du labo, qui au détour d'une conversation scientifique, d'une aide ponctuelle ou systématique, ou simplement d'une pause café, ont participé d'une manière ou d'une autre à cette thèse et à son achèvement, et qui ont rendu plus agréable le boulot : Solange, Nathalie V, Carole, Sonia, Anne, Elizabetha, Marc, Nathalie H, et les autres

Merci également aux jeunes chercheurs, dont les caractères très différents créent la bonne ambiance qui existe entre nous : Sandra, Sylvia, Marine, Krystina, Chloé, Ibrahima, et les masters ; tout ceux qui sont passés du côté des grands et ont quitté le laboratoire ; et ceux du site Ampère avec qui j'ai eu l'occasion de passer de bons moments, et qui se reconnaîtront. J'ai une pensée particulière pour les « copains de fin de rédac' » avec qui j'ai pu partager doute, espoir et raz-le-bol en sachant qu'ils me comprenaient : Maria, Paolo, Hien et Claire. Pour ceux qui n'ont pas encore soutenu : accrochez-vous et tenez bon !! Je vous envoie pleins de courage pour la fin !!

Merci à ma famille qui m'a soutenu et a été présente du début à la fin, chaque fois que j'en ai eu besoin : en particulier à Cécile et Fabrice pour leurs attentions venues de loin, et à mes parents pour leurs petits plats et leurs attentions les we et vacances de rédaction, et pour leurs corrections. Quant à la nouvelle génération, Lili, Tinoé, « kirikou » et X, j'espère pouvoir désormais passer plus de temps avec eux.

Je dédis aussi cette thèse à mes amis et autres partenaires de grimpe, ski, plongée, montagne, parapente, jujitsu, voyage, ou simplement de soirée et de bons moments. Je pense à ceux que je n'ai plus (assez) vu faute de rédac', ou ceux avec qui j'ai encore pu partager des bons moments malgré le manque de temps, la fatigue et le stress :

- les grimpeurs / skieurs / alpinistes Vio, Arnaud, Josué, Amandine, Lydia, Annaëlle et les autres de l'équipe U ;
- Fanny, Patrice, Noé et mes autres binômes d'apnée ;
- les « anciens copains », qui rentrent souvent dans plusieurs de ces catégories, Elsie, Nad, Max, et puis Marie, Colas, Sophie L, Sophie M, Axelle, Rachel ;
- les collocs et associés : Fabien, Dam, Pascal, etc. qui m'ont supporté (dans tous les sens du terme) dans la vie quotidienne et même en dehors ;
… et bien d'autres qui se reconnaîtront !

Une pensée particulière aux « coachs » et assimilés de mes activités sportives passées et présentes : Manu T, Phil P, Yannick I, Yves R, et l'infatigable Anne H, qui me permettent de me défouler, et de progresser (... ou pas !) dans une ambiance top.

Une dernière dédicace spéciale à ceux qui ont toujours été là quand il le fallait, au moins à distance. Il sont déjà cités dans ces remerciements, mais méritent une mention spéciale. Merci à ces amis pour leur présence, leur soutien, et tout ce que l'on partage : Vio, Arnaud, Marie, Josué, et puis aussi Maria, qui va la finir, cette thèse ! ;-) ).

Je terminerais en dédiant cette thèse à Nico L et ses expressions timides et subtiles, mon sujet de corpus préféré et plus encore... qui est parti beaucoup trop tôt.

# TABLE DES MATIÈRES

# LISTE DES SIGLES UTILISÉS

ACA : Agent Communicant Animés (ECA en anglais, pour « Embodied Conversational Agent »)

API : Alphabet Phonétique International

BDI: Beliefs Desire Intentions

FACS : Facial Action Coding System

FoK : Feeling of Knowing

FoT : Feeling of Thinking

IHM : Interaction Homme-Machine

# INTRODUCTION

> « Qu'il s'agisse des vestiges ou du corps d'autrui, la question est de savoir comment un objet dans l'espace peut devenir la trace parlante d'une existence, comment inversement une intention, une pensée, un projet peuvent se détacher du sujet personnel et devenir visibles hors de lui dans son corps, dans le milieu qu'il se construit. » (Merleau-Ponty, 1976, *Phénoménologie de la perception*, p. 401).

## I. Les affects dans la communication et dans les technologies

Dès l'Antiquité, la communication des émotions dans le discours représentait un intérêt certain, notamment pour la philosophie grecque, dans la rhétorique (*cf.* par exemple Ducrot & Schaeffer, 1995, p. 166-168). Cette dernière, fondée sur la tripartition *ethos*, *pathos*, *logos*[1], inspira plus tard la stylistique.

Aujourd'hui, les technologies de communication virtuelle humanisées prennent une place croissante. Il s'agit dès lors de modéliser les stratégies communicatives des humains en interaction, simuler leurs performances par un robot physique ou virtuel, et comprendre les mécanismes cognitifs qui sont impliqués. L'expression des émotions, et plus globalement des affects, est amenée à être considérée comme un processus important et indispensable à une communication écologique. Ce qui place l'humain dans le temps et le contexte de son interaction revient au centre du débat : ses motivations, ses états mentaux ou encore ses intentions au cours de l'interaction. Sous cet angle, la pertinence d'un énoncé consiste à remettre l'humain au centre de l'observation de la communication, et à observer la langue et la culture à travers sa personnalité, son rôle sociétal à l'instant de l'observation. Elle consiste également à comprendre le « sens » de ses actes de parole, comme un objet motivé et intentionnel relié au temps et au contexte de l'interaction. En somme, il s'agit de situer la communication.

De plus, au-delà du cloisonnement habituel entre cognition et affects, de nouvelles hypothèses fortes de la psychologie cognitive (*cf.* (Sander & Scherer, 2009) pour un état de l'art) et de la neuropsychologie (Damasio, 1994) donnent aux affects un statut central dans une interrelation intime avec les fonctions cognitives. À l'heure où les

[1] Plus précisément, cette tripartition, concerne selon la rhétorique : 1) l'Ethos, « Ce que je veux paraître » ; 2) le Pathos « Ce que je veux qu'ils ressentent » ; 3) le Logos, « Le langage que je vais adopter pour servir la relation entre Ethos et Pathos ». (issu de http://www.praxis-communication.com/fenetres/ethos.html, consulté pour la dernière fois le 06/03/2011)

« sciences affectives » s'imposent comme thème de recherche à part entière (l'ouvrage *Handbook on Affective Sciences* leur a été consacré (Davidson, Scherer, & Goldsmith, 2003)), les liens entre affects et cognition se retrouvent à la base des nouvelles théories des émotions (*cf.* Chapitre 1).

Ainsi, dans de récentes approches de *l'Affective Computing* (littéralement « informatique affective »), la dichotomie entre émotion et cognition d'une part, et entre linguistique et pragmatique d'autre part, commencent à s'effacer (*cf.* les actes du cycle de conférences internationales *Affective Computing and Intelligent Interaction*). Cela est du en particulier à la nécessité de modéliser des interactions écologiques, réalistes et surtout personnifiées, dans les technologies de NLP (*Natural Language Processing*) et dans le domaine de *l'Affective Computing*. En effet, les technologies visent à terme la communication face à face : les clones parlants ont besoin d'être incarnés dans un corps et un visage. Cela implique que l'expression verbale ou non-verbale est en premier lieu concernée, avec toute la complexité de la multimodalité faciale, gestuelle et parlée de ces expressions qui donne lieu actuellement à de nombreux séminaires et groupes de travail (comme par exemple le WACA -Workshop sur les Agents Conversationnels Animés - qui a lieu tous les deux ans depuis 2006 au sein de la communauté francophone). C'est aussi grâce à ces contraintes que l'intérêt s'est accru sur la prosodie, qui a alors acquis un statut important en interaction (Campbell, 2004).

Ainsi, la communication ne se réduit pas au langage : gestualité / posture, expressions faciales, interjections et autres bruits de bouche, et prosodie doivent être considérés, dans la variété et la complexité de leurs expressions. Chacune de ces modalités est susceptible d'être ou d'apporter de l'information.

## II. Le comportement expressif en question

Nous nous intéressons à tous les phénomènes perceptibles, audibles et/ou visibles, de l'expressivité, et à la manière dont ils portent une information sur la communication. En effet, nous ne nous situons pas ici dans la modélisation des processus cognitifs / affectifs qui conduisent à traiter ces expressions, mais dans la relation entre les manifestations acoustiques et / ou visuelles des effecteurs, et la perception opérée par l'humain. Nous nous demandons quels sont les traits morphologiques pertinents, et quels indices communicatifs sont véhiculés par ces morphologies. Nous notons bien entendu que ces mécanismes sont certainement imbriqués avec les mécanismes de traitement.

Des travaux préalables (en particulier Loyau (2007) et Audibert (2008)) suggèrent, en accord avec ce que la phonétique a souvent mis en évidence, que le « détail » peut être lourd en information. Notre approche va consister à considérer et décrire, sur la face comme dans la parole et la voix, les expressions mais aussi les « micro-expressions » que nous utilisons ici sous sa définition première, c'est-à-dire de « fins détails », et à poser les prémisses de leur significativité, impliquant le sujet.

Ces précédents travaux de notre équipe de recherche (*ibid*, et Aubergé (2002) et Shochi, Erickson, Rilliard, Aubergé, & Martin (2008), entre autres) ont montré que ces phénomènes ramènent toujours l'attention sur la question de la spécificité du sujet dans son interaction particulière. Les émotions ne suffisent donc pas à elles-seules couvrir le champ d'étude des expressions. Ce propos sera entre autres illustré par le Chapitre 1, qui montre que tous les auteurs sortent eux-mêmes du cadre restreint des émotions dans lequel ils s'ancraient au départ.

D'autre part, cette observation implique de commencer à étudier ces phénomènes en restreignant la situation de communication à une forme la plus simple possible.
Nous avons donc travaillé sur un corpus d'Interactions Homme-Machine (IHM), E-Wiz SoundTeacher (Aubergé, Audibert, & Rilliard, 2006), conçu à l'origine pour que les données concernent des sujets sans interlocuteur. Le but était ainsi, en utilisant le paradigme du magicien d'Oz (Chapitre 3 II.2.), de récupérer des expressions émotionnelles involontaires, mais pas d'expressions liées à l'interaction, ni d'attitudes.
Or, nous nous sommes rendus compte, grâce notamment à l'utilisation d'auto-annotations effectuées par le sujet sur ses propres états au cours de l'enregistrement (Chapitre 3 III.3.1.), que même dans ce contexte restreint d'interaction, nos sujets humains envoient des informations qui sont loin d'être réduites à des expressions involontaires. Ces informations concernent également des expressions d'états

mentaux (intentionnelles, mais aussi non-intentionnelles), ainsi que des attitudes (intentionnelles par définition) (*cf*. Loyau, 2007). De plus, nous avons pu observer que les sujets ont eu une sorte de « jugement » sur la machine, même sans avoir eu de véritable « interaction » (sociale) avec elle. Le cadre spécifique de l'IHM n'a donc pas eu tout l'effet escompté lors des enregistrement : il n'a pas permis de geler les signaux sociaux venant du sujet, même si ces derniers n'avaient pas lieu d'être.

Certains types d'éléments que nous avons observés sont également relevés dans le phénomène de *backchannel,* lors des *feedbacks* de l'auditeur (*cf.* les travaux du réseau Humaine D6d, *e.g.* Pelachaud et al., 2005). Plus globalement, ils sont produits au cours des *feedbacks* d'utilisateur / locuteur impliqué dans une tâche d'interaction homme / [homme ou machine], lors des processus cognitifs du sujet liés à la tâche. Ainsi, le sujet exprime « en ligne », et en continu dans la modalité visuelle (Loyau, 2007 ; Vanpé & Aubergé, 2010), ses états mentaux, attitudinaux ou émotionnels (ce que nous nommons *Feeling of Thinking - FoT*).

En effet, en situation d'interaction communicative (même en IHM), et y compris en dehors de ses tours de parole, un interlocuteur produit de nombreux indices et signaux : événements « biologiques » ou d'inconfort (toux, etc.) mais aussi gestes faciaux, postures, interjections et autres bruits de bouche (clicks, gémissement, etc.). Tous expriment différents types d'informations : changements d'états émotionnels, attitudes, humeurs ou états mentaux, relatifs à l'interaction elle-même, à la tâche que le locuteur est en train d'effectuer, ou plus globalement à tout paramètre situationnel de l'interaction communicative.
Nous avons choisi le terme de « *Feeling of Thinking* » *(FoT)* pour désigner l'ensemble des états exprimés, généralisant ainsi le « *Feeling of Knowing* » de Swerts & Krahmer (2005). La quantité et la nature de telles informations sont des éléments éminemment pertinents dans l'interaction. Toutefois, dans le domaine des ACAs (Agents Conversationnels Animés), si les agents sont « corporéisés » virtuellement, la gestualité du corps et de la face et autres expressions du *FoT* sont encore succinctes et ne sont pas toujours adaptées à la situation de communication. Elles sont en général réduites aux expressions émotionnelles du modèle le plus répandu : le FACS (*Facial Action Coding System* -Ekman & Friesen, 1978-). Or, ne pas les générer dans un ACA reviendrait à donner à l'interlocuteur humain des messages partiellement faux, malformés et non écologiques. En parallèle, ne pas reconnaître de telles informations en provenance de l'humain dans une interaction personne-machine priverait l'agent virtuel d'informations pertinentes.

En particulier, la gestualité de la face et du haut du buste[2] semble jouer un rôle important dans l'expression des affects, et plus précisément des émotions. Même si des recherches au sujet de l'expression des émotions par cette gestualité existent déjà, notamment en psychologie (*cf.* Chapitre 1, et Chapitre 2 I. et II.), le problème est rarement bien posé.

En effet, dans la majorité des cas, la méthodologie liée à l'essence même des corpus émotionnels et de leur annotation est mise à défaut. L'aspect le plus particulièrement remis en cause dans cette thèse concerne l'annotation : l'accent est porté sur l'importance d'une annotation « naïve » dans la mesure du possible, alors que ce sont le plus souvent des « experts » qui annotent, interprétant involontairement ce qu'ils perçoivent de par leurs capacités d'humains.

Pour remédier à cela, nous avons adopté une méthodologie inspirée de l'éthologie, qui consiste en la construction d'une sorte d'éthogramme d'interactions, avec tout ce que cela suppose en termes d'annotations et de principes à respecter. Nous traiterons en détail de tous ces aspects méthodologique dans le chapitre 3.

[2] Nous ne parlons pas ici de posture car cette dernière est finalement la conséquence de mouvements, dont l'« apex » (comme le nomme McNeill (1992) pour les mouvements), est maintenu dans le temps.

# III. Problématique adoptée, objectifs de cette thèse, et travaux réalisés

Cette thèse s'inscrit dans un projet plus global concernant l'expressivité en interaction communicative, située dans un premier temps dans l'interaction homme-machine. Elle est avant tout une proposition de méthodologie et de programme de recherche, permettant d'étudier des phénomènes et des paramètres qui semblent pertinents communicativement, et / ou important à modéliser dans un but technologique de reconnaissance ou de synthèse.

Ainsi, le travail empirique réalisé nous a seulement permis de mettre en forme et de tester cette proposition méthodologique. Il nous a également donner la possibilité de réaliser des analyses préliminaires testant certaines de nos hypothèses, et ainsi de justifier la pertinence des pistes de recherche que nous suggérons.

Notre approche est multimodale et concerne expressions faciales, gestualité du haut du buste et événements vocaux. Notre objectif global est de trouver les variables de l'expression affective (au sens large), et leurs paramètres pertinents (c'est-à-dire ce qui a un effet communicatif). Notre problématique concerne globalement l'organisation multimodale des différents aspects du *FoT* (mental, émotionnel, etc.) ainsi que l'organisation temporelle de cette expressivité. Plus particulièrement, nous nous demandons si, et de quelle manière, ces phénomènes audibles et / ou visibles sont récupérés ou non par un interactant, c'est-à-dire s'il s'agit d'indice ou de signal.
Par ailleurs, nous gardons à l'esprit que la signification de chaque événement, et plus globalement du comportement multimodal en interaction communicative, pourrait être émergent. Cela signifie que le tout ne serait pas la somme des parties, ces dernières pouvant être, dans notre cas, aussi bien les différentes modalités, que les différentes parties du visage au niveau des expressions faciales.

Par conséquent, notre approche s'attache à considérer les événements dans leur ensemble, et à ne pas partir du postulat de la séparation *a priori* des différentes modalités. Toutefois, nous avons dans un premier temps étudié des stimuli visuels pour lesquels les sujets ne produisaient pas de formes sonores, afin de réduire la complexité multimodale.

Cependant, nous ne cherchons pas à établir, comme le fait par exemple Poggi (2007), des lexiques pour chacune des modalités, mais plutôt à déterminer des paramètres pertinents pour chacune des fonctions remplies par le comportement expressif, voire

plus particulièrement des états exprimés par l'individu (ces fonctions n'étant pas listée *a priori*, de même que les paramètres potentiels).
Du point de vue des technologies, trouver quels sont les paramètres (processus dynamique -*e.g.* les travaux de Pelachaud et collègues-, rythmique, mais aussi les phénomènes eux-mêmes) qui entrent en jeu dans l'identification des états du *FoT*, peut permettre de mieux cerner ce qu'il est pertinent et nécessaire de modéliser.

Après une phase « naïve » d'étiquetage objectif des formes audibles et visibles de notre corpus (*cf.* Chapitre 3 IV. et Chapitre 5 III.), nous avons validé perceptivement certaines « icônes gestuelles », expressions et micro-expressions de la modalité visuelle, au moyen d'un test perceptif (Chapitre 4). Ce dernier a consisté à présenter les « icônes gestuelles » à des juges naïfs, dans une tâche d'association avec leur auto-annotation correspondante (*i.e.* le label que le sujet lui-même a utilisé pour décrire l'état dans lequel il se trouvait à cet instant là). Nous avons testé chaque stimulus sous une forme statique (photo) et dynamique (vidéo), et selon trois conditions de présentation (haut de visage seul, bas du visage seul, et visage entier).
Ensuite, nous avons analysé les occurrences d'événements vocaux selon leur type, en nous focalisant sur leur organisation temporelle (*i.e.* en particulier sur de potentiels motifs temporels et sur leur rythmicité), en regard de paramètres situationnels objectifs, liés au scénario et leurs paramètres (Chapitre 6). Ces analyses ont également portés sur la variabilité inter-sujets (Chapitre 6 III.).
Finalement, nous nous sommes intéressés aux caractéristiques intrinsèques des événements vocaux (Chapitre 7), et notamment au niveau de contrôle prosodique propre à chacun d'eux (en particulier par le contrôle de la durée et de la qualité de voix). Une étude perceptive menée par (Signorello, Aubergé, Vanpé, Granjon, & Audibert, 2010) a enfin cherché des indices de culture et / ou de langue parmi nos micro-événements, et ce de manière audio-visuelle.

# CHAPITRE 1 : LES PRINCIPALES THÉORIES PSYCHOLOGIQUES DES ÉMOTIONS ET LEURS DÉBATS ESSENTIELS

La pensée rationnelle est étudiée chez l'homme depuis des siècles par la philosophie et la logique. Elle a pendant longtemps été valorisée (notamment par Descartes), par opposition au registre affectif (sentiments, émotions, passions, etc.), considéré comme inhibant et auquel l'homme devait se soustraire pour fonctionner au mieux. La notion de passion était alors appréhendée comme une pathologie, et Kant, entre autres, percevait l'émotion comme un état irrationnel, une maladie de l'esprit.

À partir du 19ème siècle, des études sur les émotions menées à l'aide de méthodes expérimentales bien définies ont revalorisé le rôle du registre affectif quant à son implication lors des processus cognitifs. Elles ont montré par exemple que chez les rats, le stress permettait un temps de réaction et des actions plus rapides. Cependant, au dessus d'un certain seuil, l'inhibition due au stress était totale et empêchait toute réaction adaptée et action efficace. Les émotions sont dès lors conçues non plus comme une pathologie, mais comme le fondement du développement et du fonctionnement psychologique normal de l'individu.

L'importance des interactions entre affects et cognition n'est désormais plus niée par personne, mais théoriser ces systèmes et leurs interactions reste problématique. En effet, un grand nombre de questions sous-jacentes reste en suspens, de la définition même de l'émotion et de sa nature (un état du cerveau, un état physiologique, psychologique, un processus cognitif ?), à ce qui différencie ou lie émotions, attitudes, humeurs, sentiments ou encore états mentaux, en passant par les problèmes de catégorisation et de contrôle de tous ces états. Nous évoquerons dans ce chapitre la nature des débats et les grandes hypothèses théoriques de la psychologie décrivant les émotions (pour les approfondir se reporter par exemple au chapitre 1 du récent « Traité de psychologie des émotions » de Sander & Scherer (2009). En parallèle, nous mentionnerons également ici les implémentations de certains de ces modèles, en simulation, puisqu'elles permettent de mettre leurs performances à l'épreuve la pertinence.

# I. Un objet d'étude qui suscite des controverses

Les controverses entre les auteurs qui s'intéressent au phénomène des émotions sont liées aux ambivalences, voire à l'objectivation, portant sur l'origine et la nature même de l'émotion (*cf.* par exemple Scherer & Sangsue (1996)). Avant de devenir un objet scientifique, l'émotion était un élément du lexique consensuel de quelques cultures (celles qui se sont intéressées à scientifiser la notion d'émotion). L'ambiguïté sémantique (*e.g.* des termes émotion, sentiment, passion, et affects en français) montre qu'il n'est pas si aisée de reprendre directement des entités langagières consensuelles (issues de la conscientisation socio-culturelle des notions) en tant qu'entité scientifique objectivable. Ainsi, ces notions sont, depuis le début du siècle dernier, et encore aujourd'hui dans la langue courante, assimilées les unes aux autres ou employées les unes à la place des autres[3] :

> « Tout le monde sait ce qu'est une émotion, jusqu'à ce que vous lui demandiez de la définir. » (Fehr et Russel, 1984, cité par Christophe, 1998, p.11).

La notion d'émotion se distinguait toutefois des autres par sa connotation négative, en étant une désorganisation et une désadaptation, mais également par son intensité :

> « une forme explosive de l'affectivité qui, en envahissant le champ de la conscience, provoquait un retour aux automatismes préformés ». (Pradines, 1954, cité par Scherer & Sangsue, 1996, p.1)

Pradines *(ibid)* la qualifiait ainsi de « raté d'une régulation sentimentale » dans le comportement humain. Alors que la plupart des théoriciens actuels s'entendent sur une fonction régulatrice de l'émotion, avec un caractère adaptatif, ces caractéristiques étaient celles du sentiment pour Pradines, et avant lui pour Janet (1926).

Aujourd'hui, la communauté scientifique parle souvent d'« épisode émotionnel » pour justement marquer le caractère immédiat, soudain, temporaire et intense de l'émotion, qui apparaît en réponse à l'évaluation d'un événement, d'un stimulus, d'origine interne ou externe, mais qui « surgit » dans l'environnement du sujet. Ce sont ces caractéristiques qui la distinguent entre autres de l'humeur (comme le stress), qui est un état en général considéré comme plus diffus, durant plus longtemps, et n'étant pas nécessairement déclenché par un événement spécifique (Scherer, 2005). Quand au sentiment, il est souvent considéré aujourd'hui comme une composante de l'émotion (*cf.* partie II.3. de ce chapitre), correspondant à l'expérience consciente de l'émotion, au ressenti que l'individu a de cette dernière, qui permet à l'individu de la nommer. Il est

[3] Du moins dans le cas de langues comme le français ou l'anglais, pour lesquelles la traduction de ces notions et l'équivalence des termes n'est déjà pas immédiate

souvent plus précisément nommé « sentiment subjectif » pour éviter toute confusion avec le terme « sentiment » de la langue courante.

Il reste à préciser la distinction entre les termes « émotion » et « affect » (ou « état affectif »), qui sont souvent utilisés de manière interchangeable, comme des synonymes. Ce problème terminologique a en particulier fait l'objet de discussions lors d'une « session panel » de LREC 2006. Il en est ressorti que dans le domaine du traitement de la parole, l'expression « état affectif », plus générique que « état émotionnel », est plus adaptée pour décrire l'état « émotionnel » complexe de l'individu. En effet, cet « état affectif » inclut « les émotions, les sentiments, les attitudes, les humeurs et les attitudes interpersonnelles d'une personne » (Campbell et al., 2006, p.xxiv). À un instant T, l'état affectif d'une personne est un mélange de tous ces différents éléments, avec souvent plusieurs événements déclencheurs se produisant à différents moments. Ainsi, les états affectifs sont de nature dynamique et en changement permanent lors d'une interaction. Cela les distingue des « épisodes émotionnels » occasionnels et en réponse à un événement spécifique. Ces considérations rejoignent les suggestions de Juslin & Scherer (2005, p.69) :

> « Nous suggérons d'utiliser *affect* comme un « terme parapluie », général, qui inclut des phénomènes variés tels que l'émotion, le stress, l'humeur, les attitudes inter-personnelles et les traits affectifs de la personnalité. »[4]

Quant à nous, nous utiliserons également cette distinction terminologique entre l'expression « états affectifs » (par opposition aux « états mentaux »), et le terme « émotion », telle que suggérée par Juslin & Scherer (2005) ou Campbell et al. (2006).

Quoi qu'il en soit, tous ces phénomènes inclus dans les états affectifs, relèvent-ils d'un même phénomène psycho-cognitif ? Cette dernière interrogation reste en suspens. Il est d'autant moins facile d'y répondre qu'il est nécessaire pour cela de se soustraire aux influences des catégorisations effectuées intuitivement et dépendant probablement de paramètres socio-culturels. Les débats centraux motivant les controverses théoriques sur les émotions sont toujours le moteur principal des recherches : les émotions sont-elles de nature continue ou catégorielle ? Quelles sont leurs composantes ? Qu'est-ce qui se produit en premier, les changements physiologiques, ou le sentiment subjectif ? L'expression émotionnelle est-elle innée (et donc universelle), ou dépend-elle de facteurs socio-culturels ? Quels sont alors les universaux et les spécificités ? Nous allons tenter de retracer comment quelques réponses sont données ou discutées par différentes théories de la psychologie des émotions, dont nous donnerons un court aperçu.

[4] Citation originale : « We suggest using *affect* as a general, umbrella term that subsumes a variety of phenomena such as emotion, stress, mood, interpersonal stance, and affective personality traits. »

# II. L'émotion en psychologie cognitive

## *II.1. La perspective cognitive des émotions*

Les années 1960 ont vu la naissance des sciences cognitives, qui se sont préoccupées, en ce qui concerne l'étude des émotions, à établir les relations entre ces dernières et les processus cognitifs. Un autre point commun entre les théories du courant cognitif est leur approche fonctionnaliste des émotions, au sens où elles leur attribuent au moins les fonctions expressives et communicatives (cf. les différentes considérations quant aux fonctions émotionnelles partie IV. de ce chapitre). Cependant, leurs différences sont marquées par une plus ou moins grande prise en compte du contexte social ainsi que par les autres fonctions qui leur sont attribuées.

Dans la perspective cognitive, l'émotion est considérée comme « l'une des réalisations principales de l'évolution ». En effet, la séquence Stimulus-Réponse demande seulement un traitement rudimentaire, automatique et inné des stimuli (Lorenz, 1965). Elle est ici remplacée par un « mécanisme hautement flexible qui "découple" le stimulus et la réponse » (Scherer & Sangsue, 1996, p.4).

## *II.2.* Le concept d'*appraisal* d'Arnold puis Lazarus, et la théorie de Frijda

Arnold (1960) a été l'une des premières à suggérer que le cerveau est très actif dans le décodage de stimuli émotionnels. Dès les années 1950, elle a décrit les émotions comme fondée sur l'*appraisal*, processus cognitif qui permet d'apprécier les événements environnementaux. Le principe est qu'une évaluation positive d'un stimulus déclencherait une émotion positive et un comportement de rapprochement. À l'inverse, une évaluation négative produirait une émotion négative et un comportement d'éloignement.

La théorie d'Arnold a ainsi réintroduit le rôle de l'environnement social dans la façon de percevoir et d'évaluer cognitivement le comportement physiologique. L'émotion ensuite ressentie serait déterminée par la signification que nous attribuons à la situation dans laquelle nous nous trouvons, signification qui nous est propre et qui est influencée par nos expériences passées. C'est ce qui est appelé « évaluation subjective », et qui est un processus « direct, immédiat et intuitif, et peut être conscient ou inconscient » (Nugier, 2009, p.10). Ce concept d'*appraisal*, central dans les théories cognitives des émotions développées par la suite, expliquerait donc qu'un même événement puisse déclencher des émotions différentes selon les individus ou même

chez un même individu à des moments différents. Ces différentes étapes du processus émotionnel peuvent être résumées par ce schéma :

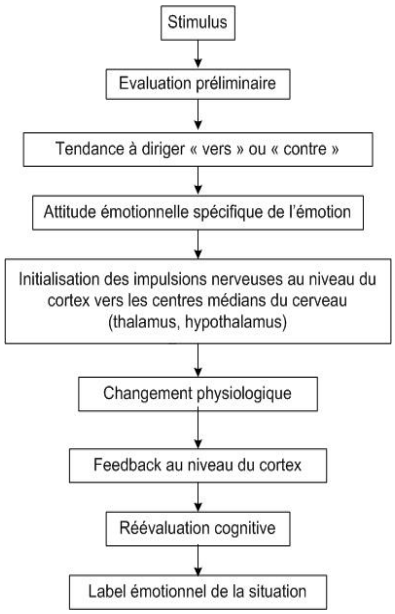



Figure 1: Représentation de la théorie d'Arnold (1960), d'après Christophe (1998, p.40).

Chaque expérience émotionnelle correspondrait à un patron particulier d'évaluations cognitives (l'*appraisal*) sur différents éléments situationnels et avec une influence de facteurs sociaux et environnementaux.

De plus, Arnold a introduit dans son modèle le concept de « tendance à l'action » qui sera repris et développé dans des études postérieures, et notamment celles de Frijda.

Lazarus (1991) reprend la notion d'*appraisal* à la suite d'Arnold. Il insiste sur le fait qu'une même situation déclenche des émotions différentes selon l'interprétation que l'individu en fait :

> « L'individu éveillé évalue en permanence sa relation à l'environnement, et ce relativement aux implications que celle-ci peut avoir pour son bien-être personnel. » (cité par Adam & Evrard, 2005, p. 136)

C'est en s'inspirant de ce courant de psychologie qu'a été mis au point le modèle computationnel des émotions d'Ortony, Clore, & Collins (1988), dit OCC. Bien que très réductionniste, ce modèle est très utilisé dans le domaine des IHM, des agents conversationnels ou encore des robots humanoïdes, par sa praticité d'implémentation. Il « calcule l'émotion » qu'un locuteur va ressentir et / ou exprimer, à partir d'un ensemble de règles logiques et de conditions. Ainsi, « dans ce modèle les émotions sont déclenchées par l'évaluation de changement dans le monde, changement pouvant venir d'un objet, d'un événement ou d'une action. Suivant les résultats de ces évaluations, […] une certaine émotion sera déclenchée. » (Pelachaud, 2006 p. 442).

Il est à noter que le modèle OCC s'inscrit également dans le domaine des agents autonomes dits BDI (pour *Belief Desire Intention*). Ce sont des agents décrits par leurs états de croyances, désirs, et intentions. Tout comme dans le modèle OCC, les états affectifs sont considérés dans ce domaine des agents BDI comme issus d'une logique, le corps étant alors une commande extérieure et une sortie, et non un système de contrôles et contraintes.

Revenons à Frijda (1986). Les émotions sont pour lui des états de motivation qui incitent les individus à agir et à s'exprimer. Ces états seraient issus d'une évaluation de l'environnement, et se révèleraient à travers des états de préparation à l'action. Les actions résultantes du sujet peuvent lui servir à établir, maintenir ou rompre l'interaction avec autrui, ou bien à modifier sa relation à l'environnement. Frijda en relève 18, qu'il regroupe en six catégories (Frijda, 1987) : « vers », « contre », « activation », « désactivation, évitement et inhibition » et « interruption ».

Cette tendance à l'action accompagnerait l'expérience émotionnelle de manière spécifique à un individu particulier dans une situation donnée, même si elle reste typique à chaque émotion. Par exemple *vouloir s'enfuir* serait typique de la peur, alors que *vouloir attaquer* serait typique de la colère.

### *II.3. La théorie de Scherer*

Scherer a développé un modèle, dit « des processus composants », sur lequel se fondent de nombreuses recherches actuelles. Il montre aussi un intérêt particulier pour l'expression émotionnelle à travers la parole.

En se fondant sur une approche phylogénétique[5] et comparative, il décrit les émotions comme des agents intermédiaires entre un environnement changeant et l'individu. Ainsi, les émotions rempliraient essentiellement des fonctions adaptatives. Toutefois, elles ne seraient cependant pas assimilables à des réflexes car la réponse résulte, comme pour Arnold, des évaluations cognitives du stimulus (Scherer, 1984).

Dans la lignée de Lazarus (1991), Frijda et Scherer attribuent aux processus émotionnels trois autres fonctions majeures :

- l'évaluation de la signification des stimulations par rapport aux besoins, projets ou préférences d'un organisme ;
- la préparation physiologique et psychologique aux actions propres à répondre à ces stimulations du milieu ;
- la communication des états et intentions de l'organisme à son environnement social, puisque les émotions sont rendues visibles par le biais des expressions motrices et de

[5]La phylogenèse est le développement des espèces vivantes au cours des temps, par opposition à l'ontogenèse qui se limite au développement de l'individu, depuis la fécondation de l'œuf jusqu'à l'âge adulte.

la préparation aux actions. C'est en ce sens qu'elles sont capitales pour le développement de la vie sociale.

Les émotions sont considérées par Scherer comme multidimensionnelles. Dans son *component process model* (modèle des processus composants), Scherer (1984) envisage l'émotion comme une série de changements de l'organisme qui s'adapte à l'environnement au fur et à mesure des processus d'évaluation des stimuli. Les émotions y sont décrites en termes de processus composants et de séquences successives d'évaluation en réponse à un stimulus externe ou interne. Les émotions sont donc envisagées comme un processus multidimensionnel à cinq composantes, qui ont chacune leur propre fonction dans le système émotionnel (Scherer, 1999) :

- la composante cognitive : évaluation de l'environnement, des stimuli ;
- la composante physiologique : fonction de régulation ;
- la composante de l'expression motrice : communication des intentions ;
- la composante motivationnelle : préparation aux actions ;
- la composante de sentiment subjectif : réflexion et contrôle.

Toujours selon ce modèle, différents critères d'évaluation interviennent dans l'expérience émotionnelle. Le traitement du stimulus passe par cinq étapes successives, les « S*timulus Evaluation Checks* » *(SEC's*), toujours dans le même ordre temporel, de la plus simple à la plus complexe :

- l'évaluation cognitive de la nouveauté ou du caractère inattendu de l'événement, c'est-à-dire s'il s'est produit ou non un changement dans les stimuli internes et externes ;
- l'évaluation du caractère intrinsèque plaisant ou déplaisant du stimulus, de laquelle découle une tendance au rapprochement ou au recul de l'organisme ;
- l'évaluation des buts et intérêts favorables, c'est-à-dire la manière dont le stimulus avantage ou entrave l'atteinte des buts.
- l'évaluation de la capacité de maîtrise de l'individu face à une situation négative (selon cette dernière, par exemple, l'émotion engendrée sera la colère si la situation n'est pas jugée dangereuse, et la peur ou la dépression à l'inverse) ;
- l'évaluation de compatibilité avec les normes et l'image de soi (*e.g.* il y a apparition de la honte si cette évaluation indique que le comportement de l'individu n'est pas conforme aux normes).

Outre son intérêt pour les mécanismes cognitifs qui entrent en jeu lors du processus émotionnel, Scherer a également porté particulièrement son attention sur l'expression des émotions, notamment dans la parole.

À ce propos, il fait une distinction entre des facteurs « push » et « pull » de l'expression (Scherer & Sangsue, 1996). Les premiers « poussent » l'expression affective, en particulier via les conséquences physiologiques. Ils concernent « l'externalisation biologiquement déterminée de processus internes de l'organisme apparaissant naturellement » (Scherer, 1994, p.183). Les « pull effects » « tirent » quant à eux l'expression à travers les modèles socialement médiatisés auxquels nous sommes soumis. Ils concernent « les normes ou les moules socio-culturellement déterminés, concernant les caractéristiques du signal requis par les codes socialement partagés de la communication des états mentaux et des intentions comportementales » (*ibid).*

Nous reviendrons sur les travaux de Scherer concernant l'expression vocale des émotions au cours des Chapitre 5 et 7.

## *II.4. La perspective cognitive actuelle et son approche componentielle de l'émotion*

Depuis les années 1980, les travaux menés au sein de cette perspective cognitive des émotions cherchent en particulier à préciser et étoffer les dimensions prises en compte lors du processus d'évaluation. Malgré le fait que des différences existent encore entre les modèles, concernant la nature et le nombre des dimensions qui interviennent lors du processus, de fortes similarités apparaissent. Ainsi, les dimensions d'évaluation les plus largement incluses dans les différents modèles sont au nombre de quatre (Nugier, 2009), mais elles peuvent se subdiviser en sous-dimensions (voir Grandjean & Scherer, 2009, pour plus de précisions) :

- la « détection de la pertinence »[6] et l'« évaluation de l'implication », qui sont souvent évaluées de façon automatique et inconsciente (Scherer, 1984), et qui déterminent le niveau d'attention alloué à l'événement déclencheur ;
- le « potentiel de maîtrise » et l'« évaluation de la signification normative », qui nécessitent un traitement cognitif plus complexe.

Nous retrouvons donc les valeurs de *SEC's* qui étaient présentes dans le modèle de Scherer (*cf.* partie II.3.). Selon les auteurs, la séquence de ces évaluations est considérée comme fixe ou flexible. Dans ce dernier cas, la séquence est déterminée par le stimulus et des facteurs environnementaux.

[6] La notion de pertinence est liée à la théorie de Sperber & Wilson (1989) : la « théorie de la pertinence ». Cette dernière est fondée sur le principe que les humains sont faits pour chercher de l'information pertinente, la pertinence de l'information étant définie en termes d'effets cognitifs gagnés et d'efforts produits pour traiter l'information : plus les effets cognitifs gagnés sont grands et les efforts de traitement faibles pour acquérir ces effets, meilleure est la pertinence de l'information.

> « L'idée générale est que l'évaluation subjective d'une situation, d'un événement ou d'un stimulus, sur la base de ces critères d'évaluation, détermine la nature (c'est-à-dire à la fois la qualité et l'intensité) de la réaction émotionnelle. » (Scherer & Sangsue, 1996, p.6)

L'émotion est ainsi envisagée comme résultante de la perception et de l'évaluation d'un événement ou d'un stimulus, externe ou interne, et elle se traduit par des comportements expressifs. La nature même de l'émotion est déterminée par l'évaluation cognitive.

Aujourd'hui, les différentes théories s'accordent pour qualifier l'émotion de « phénomène multicomponentiel adaptatif » caractérisé par trois composantes principales, pour lesquelles Scherer a fortement contribué à démontrer la pertinence expressive, non seulement de la face et des gestes, mais aussi de la parole :

- la réponse psycho-physiologique (*e.g.* la fréquence cardiaque, le flux sanguin et la production des larmes) ;
- l'expression motrice (du visage, de la voix et des gestes) ;
- le sentiment subjectif, c'est-à-dire ce qu'on pense ou dit ressentir, le terme « sentiment » étant donc à différencier du terme « émotion ».

Au-delà de cette « triade de la réaction émotionnelle » (Sander & Scherer, 2009), s'ajoutent le plus souvent deux autres composantes essentielles :

- la tendance à l'action (voir Frijda, partie II.2. de ce chapitre) ;
- la composante d'évaluation cognitive (*appraisal*), qui détermine les changements dans les quatre autres composantes.

Dans le cadre des théories cognitives, ce type de modèle, fondé en particulier sur les recherches de Scherer, a été implémenté afin de générer les expressions faciales liées aux émotions (*e.g.* Kaiser & Wherle, 2006). Un grand nombre de recherches ont été menées et il s'avère que les prédictions de ces dernières sont relativement valides (*cf.* entre autres Chapitre 2 I.3.) et généralisables à un grand nombre de situations. C'est pourquoi nous allons adopter cette conception de l'émotion, en tant que phénomène multicomponentiel.

Nous allons maintenant préciser les principaux débats concernant la composante physiologique de l'émotion, puis ses fonctions adaptatives et sociales (en particulier communicative), à travers un aperçu des autres perspectives adoptées pour l'étude des émotions.

# III. La perspective physiologique

Les théories physiologiques, dont les principales théories d'origine sont celles de James, Lange et Cannon, se sont intéressées au rôle de l'activation physiologique dans le déclenchement des processus émotionnels et ont dominé les recherches pendant de nombreuses années.

## *III.1. La théorie James-Lange, ou théorie périphérique des émotions*

Dans son ouvrage « The principle of psychology » (1890) et son article « What is an emotion ? » (1884), Williams James pose la question de la nature de l'émotion d'un point de vue physiologique. Son idée -soutenue quasi simultanément par Carl Lange (1885)- est que faire l'expérience d'une émotion, c'est d'abord faire l'expérience des changements corporels ou physiologiques qui l'accompagnent. Sans la perception de ces changements, il est impossible de faire l'expérience de ses émotions. Cette théorie a souvent été résumée par la phrase : « On se sent triste parce qu'on pleure. »

James et Lange donnent une séquence identique dans le processus émotionnel (représenté Figure 2) en occultant tous deux l'aspect social des émotions. C'est pourquoi leurs théories sont fréquemment associées sous le nom de « théorie James-Lange ».

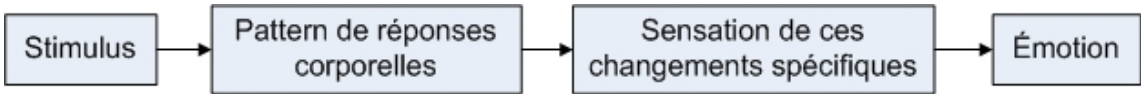



Figure 2: Représentation schématique de la théorie James-Lange (1984-1985)

L'émotion serait donc moins rapprochée de sa cause (le stimulus) que de ses effets (les changements corporels).

Cependant James et Lange sont en désaccord sur un point :

- pour Lange, il existerait des patrons différenciés et spécifiques de réponses pour chaque émotion, et ces changements vasculaires seraient contrôlés par un centre spécifique dans le cerveau, qu'il nomme « centre vasomoteur » ;

- pour James, ce seraient plusieurs centres sensori-moteurs qui détermineraient l'émotion. Les processus émotionnels seraient donc liés exclusivement au corps, en ignorant les processus mentaux d'évaluation de la situation. Plus tard, Damasio critiqua cette dernière idée qui n'autorise aucun rôle de l'émotion dans les processus cognitifs. De plus, contrairement à Lange, James croit que des combinaisons

d'activations des organes différentes peuvent parfois sembler à peu près identiques, et ainsi renvoyer à la nomination d'un état subjectif identique.

James illustre cette idée à l'aide de la métaphore du corps comme instrument de musique (Philippot, 2004, p.38), dont les sons qui en émanent (les sentiments subjectifs) seraient le produit des différents accords joués par les cordes (combinaisons d'activations différentes des organes du corps), certains accords différents pouvant sembler similaires et être catégorisés sous une même étiquette.

Cette théorie James-Lange a fait l'objet d'une lourde controverse instiguée par Cannon dans les années 1920. Comme James et Lange, Cannon croit à l'existence d'un lien fort entre changements physiologiques et expériences émotionnelles subjectives, mais pour lui, c'est le système nerveux central qui est à l'origine de l'émotion (Cannon, 1927).

Cette « théorie centrale des émotions » repose sur diverses observations, telles la persistance d'une réaction émotionnelle chez les animaux, même lorsque les réponses végétatives sont rendues impossibles, l'absence d'induction émotionnelle lorsque les réponses végétatives sont reproduites artificiellement, ou encore le fait que des changements viscéraux apparaissent dans des états émotionnels différents et dans des états non émotionnels (*e.g.* la fièvre ou la digestion), ce qui remet en cause l'idée de James selon laquelle chaque émotion spécifique est causée par un patron d'activations corporelles particulier.

Toutefois, les critiques de Cannon à l'égard de la théorie James-Lange ne font pas l'unanimité, et auront finalement surtout permis à la théorie périphérique de se développer. Le débat de la séquence temporelle et du lien causatif entre réactions corporelles et émotion, ou plutôt sentiment subjectif, reste quant à lui encore d'actualité aujourd'hui, malgré l'apport des théories cognitives des émotions (*cf.* partie II.4.) qui a tenté de le résoudre par sa définition componentielle de l'émotion.

### *III.2. Le développement de la théorie de James et son influence contemporaine*

James a développé par la suite sa pensée et a fondé en 1892 les bases des théories de la rétroaction corporelle (*proprioceptive feedback theories*), qui mettent en avant l'aspect rétroactif de l'expression émotionnelle sur le ressenti :

> « Si notre théorie est vraie, elle devrait avoir pour corollaire nécessaire que toute évocation volontaire et dépassionnée de ce que l'on croit être les manifestations d'une émotion particulière devrait nous procurer cette émotion elle-même. » (James, 1892, cité par Sander & Scherer, 2009, p.5)

Dans ce contexte, des études empiriques se sont particulièrement intéressées à l'expression faciale et ont fondé l'hypothèse de rétroaction faciale (*Facial Feedback Hypothesis*). Cette hypothèse, fondée sur le postulat de Tomkins (1984) selon lequel l'expression du visage joue un rôle central dans la régulation émotionnelle, estime que les mouvements faciaux modulent le ressenti émotionnel.

D'une façon générale, les recherches ont montré que les variations de l'expression faciale sont corrélées positivement avec les variations du ressenti émotionnel (*cf.* Soussignan, 2002). Par exemple dans une des études classiques (Lanzetta, Cartwright-Smith, & Eleck, 1976), il a été demandé à des participants de soit amplifier, soit supprimer leur expression faciale alors qu'ils recevaient des chocs électriques. Le prétexte était qu'ils devaient induire en erreur des observateurs. Les résultats font apparaître que les participants qui ont tenté de supprimer leur expression ont jugé les chocs électriques moins douloureux que ceux qui l'ont amplifiée.

Un article d'Ekman, Levenson, & Friesen (1983) est même allé jusqu'à suggérer que l'induction d'expressions motrices particulières pouvait non seulement amplifier l'émotion mais aussi la déclencher physiologiquement. Toutefois, cette étude a été largement critiquée et débattue en raison d'artefacts et de biais expérimentaux.

La perspective Jamesienne a également influencé des études concernant les « théories de la cognition incarnée » (*embodied cognition*) et les « théories de l'esprit incarné » (*embodiement*). D'une part, ces théories cherchent à expliquer certains phénomènes des relations interpersonnelles, comme celui de l'empathie. Par exemple Niedenthal (2007) a mis en évidence que l'imitation (volontaire ou non) de l'expression faciale d'une personne faciliterait la compréhension de ce que ressent la personne imitée. D'autre part, elles s'intéressent aux représentations cérébrales des processus corporels impliqués lors de l'acquisition d'une connaissance et suggèrent que ces processus seraient à nouveau instanciés lors de la récupération mnésique de cette connaissance. Ainsi, la perception et la pensée liées à une émotion impliquent chez l'individu une « "ré-expérience" perceptive, somatosensorielle et motrice de l'émotion correspondante »[7] (Niedenthal, 2007, abstract). Comme Sander et Scherer l'ont relevé, « dans les théories contemporaines, l'importance est surtout donnée à la représentation cérébrale de l'activité somatique plutôt qu'à l'activité somatique elle-même. » (Sander & Scherer, 2009, p.7).

[7] Citation originale : « perceptual, somatovisceral, and motoric reexperiencing [...] of the relevant emotion in one's self. »

## *III.3. L'approche neurophysiologique de Damasio*

Revenons à l'idée néo-jamésienne du rôle causal des changements corporels dans l'émotion. Cette idée reste encore défendue aujourd'hui, notamment par la théorie des marqueurs somatiques de Damasio (1994). Selon lui, les événements de l'environnement (ou la représentation mentale d'événements) peuvent être marqués par l'activité somatique qui a lieu au moment de leur traitement. Cela implique que lorsque la personne se retrouve face à un événement similaire, les marqueurs somatiques peuvent être à nouveau activés, ce qui peut alors influencer les prises de décisions liées à l'événement en question.

Ce lien entre les processus émotionnels et la raison, par le biais de la prise de décision et plus globalement du comportement, constitue le fondement des recherches de Damasio. Ce dernier est en particulier reconnu pour avoir mis un terme à la dichotomie entre émotion et raison. Le titre de son livre *L'erreur de Descartes, la raison des émotions* (1994), illustre bien sa position.

Pour aboutir à ce constat, Damasio s'est tout d'abord penché sur le célèbre cas de Phineas Gage : en 1848, à la suite d'une explosion, cette personne s'est fait traverser le crâne par une barre de fer. Après cet accident, alors qu'en apparence il a gardé toutes ses facultés (marche, parole, etc.), son médecin ainsi que son entourage ont rapporté un changement radical de comportement. Auparavant responsable et très apprécié, Gage est devenu grossier et incapable de prendre une décision.

Grâce aux progrès techniques, Damasio a pu par la suite déterminer la région du cerveau endommagée et en a déduit que le changement de personnalité de Gage en était une conséquence directe. Le cortex préfrontal ventro-médian serait donc le siège des processus de prise de décisions mais aussi des processus émotionnels. À partir de là, il a conclu que la capacité à exprimer et ressentir des émotions était indispensable à la mise en œuvre des comportements rationnels (Damasio, 1994).

Ce lien entre processus neuraux et changements somatiques apparaît dans la conception des émotions par Damasio : pour lui, les émotions sont le fruit de la combinaison des processus d'évaluation mentale d'une situation réelle ou imaginaire (processus de représentation) avec des réponses corporelles à ces représentations. L'état corporel est alors signalé en retour au cerveau. Ainsi, c'est la juxtaposition de l'image corporelle et de l'image de la situation réelle ou imaginaire qui constitue l'émotion.

Damasio a donc introduit la notion d'« image mentale », qui expliquerait la possibilité de simuler dans le cerveau la perception d'un état corporel inexistant, et en conséquence des états corporels émotionnels simplifiés, grâce à des mécanismes

neuraux. Ces images mentales seraient acquises par l'association répétée de certaines situations et des états corporels survenus lors de ces situations.

La Figure 3 représente schématiquement la différence entre les mécanismes de perception d'émotions réelles *vs.* simulées :

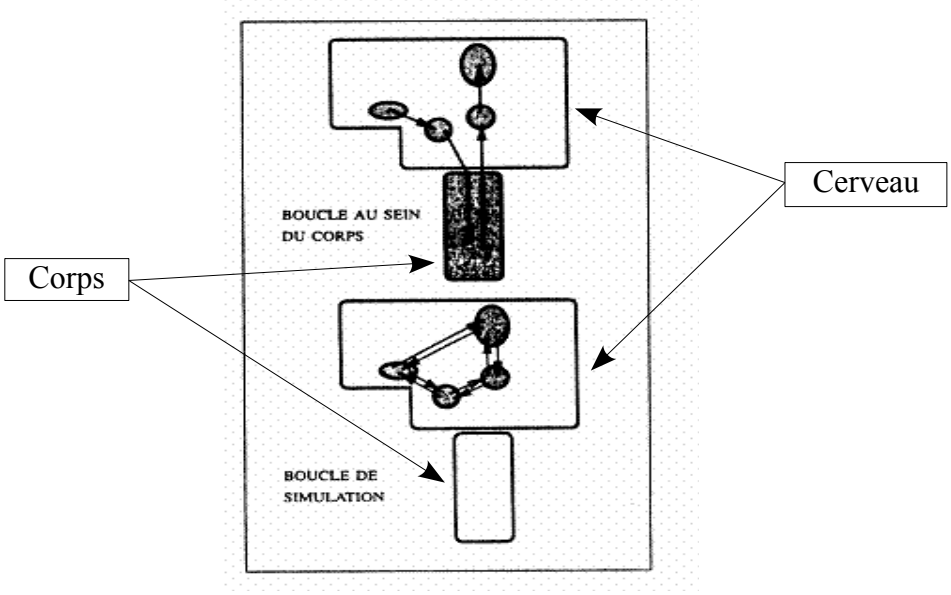



Figure 3: Mécanismes de perception des émotions, en boucle au sein du corps ou par le biais d'une boucle de simulation (Damasio, 1994, p.216)

Dans le cas des émotions authentiques, les mécanismes de perception se déroulent au sein du corps et du cerveau, alors que dans le cas des émotions simulées, la boucle s'effectue uniquement au niveau du cerveau et court-circuite le corps.

Les émotions spontanées et simulées présentent donc des différences au niveau même de leur fonctionnement cognitif.

# IV. Les fonctions adaptatives et sociales des émotions

## *IV.1. Darwin, ou des émotions universelles à fonction adaptative*

Dans la théorie physiologique (*cf.* partie III.), les changements corporels sont aussi considérés comme des réflexes génétiquement préprogrammés. Ces changements, associés à chaque émotion rempliraient ainsi une fonction de survie. Cette conception des choses n'était pas nouvelle et rappelle l'idée darwinienne de l'adaptation à l'environnement extérieur par des réponses émotionnelles automatiques.

Dans son fameux ouvrage sur l'expression des émotions chez l'homme et chez l'animal (*The Expression of Emotions in Man and Animals*), Darwin (1872) posa les premiers postulats qui influenceront les recherches sur les émotions (voir Nugier (2009) pour plus de détails) :

- les émotions sont universelles, c'est-à-dire qu'il est possible de les trouver dans toutes les cultures et dans tous les pays ;
- les émotions sont adaptatives, c'est-à-dire qu'elles auraient favorisé la survie de l'espèce en permettant aux individus de répondre de façon appropriée aux exigences environnementales.

Cette perspective darwinienne se centre en majeure partie sur les fonctions des émotions dans le contexte de la sélection naturelle, en particulier par le bais des expressions faciales. L'exemple le plus imagé est celui de l'expression faciale relative au dégoût. Puisque la fonction du dégoût était dans le passé d'éviter d'ingérer des substances dangereuses, son expression est associée à l'expulsion de nourriture par la bouche. Malgré l'évolution et le risque d'intoxication désormais minime, l'expression faciale est restée. Dans la même idée, lever les sourcils améliorerait l'acuité visuelle et retrousser le nez permettrait d'atténuer la perception d'odeurs désagréables. Ainsi les expressions faciales émotionnelles auraient acquis leurs configurations actuelles par l'interaction avec l'environnement physique. Des primitives d'expressions faciales ont d'ailleurs été retrouvées chez des animaux lors de recherches en éthologie (Van Hooff, 1972 ; Redican, 1982 ; Scherer, 1985).

Darwin a également essayé de démontrer l'universalité des émotions par le biais de l'étude de leurs expressions faciales par des observateurs, dans des peuples issus de lieux et de culture variés : d'Afrique, d'Amérique, d'Australie, à Bornéo, en Chine, en Inde, en Malaisie et en Nouvelle Zélande. Ces études l'amenèrent à déclarer que le

même état d'esprit était exprimé dans le monde entier avec une uniformité remarquable (Darwin, 1872.

Toutefois sa méthodologie est critiquable sur plusieurs points (Ekman, 1999) : le nombre de sujets était insuffisant pour une validité statistique ; Darwin s'est fondé sur les impressions données par ses observateurs anglais plutôt que de faire appel à des sujets natifs ; la formulation des questions donnaient souvent implicitement la réponse attendue (*e.g.* « La surprise est-elle communiquée par les yeux et la bouche grands ouverts, et par les sourcils levés ? »). Malgré ces biais expérimentaux, les travaux de Darwin ont eu une influence considérable sur les études ultérieures des émotions.

## *IV.2. De l'universalité des émotions à la théorie des émotions de base*

Concernant cet aspect universel des émotions, beaucoup de chercheurs ont cherché depuis les années 1960 à démontrer l'universalité de certaines expressions émotionnelles. Des expériences similaires à celles de Darwin ont été menées, en particulier sous l'impulsion de Sylvan Tomkins (1980), puis Ekman (Ekman & Friesen, 1978 ; Ekman, 1989, 1999) et Carroll Izard (Izard & Ackerman, 2000). Leur hypothèse était qu'il existe un certain nombre d'émotions particulières dont les expressions faciales sont universellement partagées et spécifiques à chacune de ces émotions.

Dans la plupart des cas ces études ont montré qu'une majorité de sujets, bien que de cultures différentes, reconnaissait la même émotion pour une expression donnée (voir des exemples d'expression faciale figure 4).


Figure 4: Exemples de photographies utilisées par Ekman dans le cadre d'études interculturelles sur les expressions faciales (Kaiser, Wehrle, & Schenkel, 2009, p.88).

Afin d'éliminer certains biais méthodologiques et de répondre aux critiques qui avaient été formulées, Ekman a effectué certaines expériences dans des communautés n'ayant pas accès aux médias, notamment en Nouvelle-Guinée, pour éviter toute forme d'apprentissage. Les résultats obtenus ont été similaires. Il a également mené par la suite des études sur les expressions spontanées de sujets américains et japonais, induites par la projection de films. Il s'avéra que lorsque le sujet était seul, ses expressions faciales étaient les mêmes quelle que soit sa culture, mais une différence entre japonais et américains apparaissait en présence d'un autre sujet de même nationalité : alors que le sujet américain ne modifiait pas son comportement, le sujet japonais masquait ses émotions par des sourires polis. Il existerait donc une différence culturelle en termes de « règles d'expressions » présentes ou non selon les cultures.

Ekman a conclu de toutes ses études que l'expression des émotions est universelle, mais qu'« en revanche, ce qui peut varier en fonction des cultures ce sont les règles d'expression des émotions, (…) ou encore les conditions de déclenchement de telle ou telle émotion » (Ekman, 1989, p.198).

Pour les néo-Darwinistes, l'universalité va de pair avec l'existence d'un petit nombre d'émotions considérées comme discrètes, notamment car elles représenteraient des patrons hautement différenciés de réponses spécifiques (physiologiques et comportementales), qui seraient génétiquement programmés et directement reliés à la survie de l'espèce (Shaver, Schwartz, Kirsona, & O'Connor, 1987). Ces émotions particulières, en nombre restreint, sont appelées *émotions de base* (Ekman), *émotions primaires* (Plutchik et Tomkins), ou encore *émotions fondamentales* (Izard).

À partir de l'étude des expressions faciales mais aussi d'observations physiologiques, Ekman (1989) a proposé un ensemble de critères pour identifier une émotion de base parmi l'ensemble des manifestations affectives. Elle doit :

- se déclencher rapidement et être brève ;
- se retrouver chez les primates non-humains ;
- être universellement identifiable (*e.g.* à partir des expressions faciales) ;
- être associée à une physiologie propre (réponse nerveuse, rythme cardiaque, etc.) ;
- se déclencher automatiquement.

Il a ainsi identifié six émotions de base : joie, surprise, peur, dégoût, colère et tristesse. Ces six émotions de base sont les plus fréquemment citées dans la littérature et sont couramment appelées *Big Six* (Cornelius, 2000).

## *IV.3. Débats et controverses liées aux émotions de base*

Les avis sont toutefois partagés parmi les tenants de la perspective néo-Darwinienne, quant au nombre et aux catégories d'émotions de base, d'autant plus que certains critères d'inclusion sont variables selon les chercheurs : neuf émotions primaires pour Tomkins (1980), auxquels il fait correspondre neuf expressions faciales primaires ; huit pour Plutchik (1970), chacune liée à une fonction adaptative distincte ; ou encore neuf émotions fondamentales pour Izard & Ackerman (2000), dont ils cherchent le rôle dans le développement de la personnalité.

De plus, l'existence même des émotions de base est sujet à controverses. L'une des raisons tient au critère de l'universalité utilisé pour déterminer ces émotions, qui ne fait pas l'unanimité. Pour les opposants à l'existence de ces émotions de base, qui appartiennent surtout au courant du constructivisme social (Averill, 1980) (*cf.* partie IV.4.), les différences culturelles influenceraient la construction de catégories universelles.

Le débat concernant l'innéisme des émotions de base reste également en suspens. Il est lié à la question de l'universalité puisque par définition, tout ce qui est inné est universel, même si la réciproque n'est pas forcément vraie. Au niveau de l'expression émotionnelle, l'idée aujourd'hui la plus répandue est qu'elle serait innée, et que nous, humains, apprenons seulement, étant enfant, les contextes dans lesquels les ressentir et les exprimer.

Les émotions seraient donc authentiques et spontanées. Par contre, la possibilité de les inhiber et leur contrôle volontaire seraient deux compétences acquises, qui pourraient être exercées. Par exemple lorsque que nous avons un revolver pointé sur nous, un mécanisme attentionnel déclenchant la peur se produit de manière automatique, mais certains savent inhiber cette peur et ne pas fuir.

Par ailleurs, les états émotionnels sont toujours nommés par des catégories, d'ailleurs souvent sans correspondance entre les différentes langues. Pourtant, de nombreux chercheurs pensent que nous avons une perception catégorielle des émotions à partir d'un continuum, de la même manière que nous percevons les couleurs. Déjà les philosophes grecs se divisaient entre partisans d'une perception catégorielle des émotions et partisans d'un continuum d'émotions différenciées selon des dimensions.

Pour la théorie « des émotions discrètes », toute émotion « complexe », dite « secondaire » par Plutchik (1984), dériverait d'une combinaison des émotions primaires (voir un exemple d'après Plutchik figure 5).

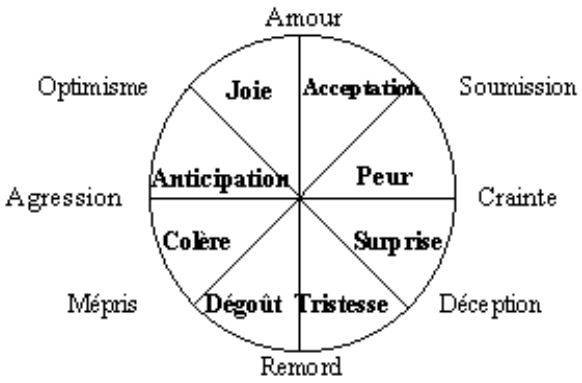



Figure 5: Modèle circulaire des huit émotions primaires (en gras) et de leur mélange, d'après Plutchik (1984)

Scherer (1984) a formulé deux critiques à l'égard de cette théorie, qu'il a nommée « théorie de la palette » par analogie aux couleurs fondamentales qui peuvent se mélanger à l'infini sur la palette du peintre :

- d'une part aucune démonstration ne prouve l'existence d'un nombre réduit d'émotions biologiquement déterminées, c'est-à-dire avec un patron d'excitations nerveuses spécifique à chacune de ces émotions ;
- d'autre part, selon lui, si des émotions opposées pouvaient se mélanger et n'en former plus qu'une, cela signifierait qu'un événement pourrait déclencher simultanément des émotions aux fonctions adaptatives complètement différentes.

D'autant plus, qu'aucun auteur n'a encore défini la proportion de ces mélanges pour aboutir par exemple aux 550 états émotionnels désignés par les adjectifs anglais (Averill, 1975).

Concernant plus précisément les expressions faciales, Sander et Scherer remettent également en question l'utilisation d'expressions actées lors des études concernant les émotions de base et leur universalité (Sander & Scherer, 2009, p.30 à 32). Ces expressions jouées par des acteurs sont certes bien identifiées lorsqu'elles sont données à percevoir à des personnes, et cela de manière interculturelle, mais elles diffèrent des expressions émotionnelles naturelles qui sont produites spontanément. Nous reviendrons sur cette problématique des expressions dans le chapitre 2, en particulier partie II.3.2., et au cours de la partie II. du chapitre 3, concernant les implications méthodologiques de cette réflexion.

## *IV.4. Le rôle des émotions dans la régulation des interactions sociales*

Un dernier apport important de Darwin a ouvert une perspective plus sociale aux théories des émotions : il suggéra l'existence d'une *fonction communicative* (ou *de signal*) de l'expression faciale. Sa raison d'être serait qu'elle permettrait aux individus d'une

même espèce d'être informés de ce que ressentent leurs congénères, et des actions qu'ils sont susceptibles d'entreprendre dans certaines situations.

Dans sa lignée, Wallon (1934 ; 1938), puis Malrieu (1952) ont étudié l'émotion et sa nature sociale en tant que nécessité adaptative. En effet, pour eux, la fonctionnalité des émotions s'exprime au travers de leur action sur autrui. De plus, le phénomène émotionnel, c'est-à-dire la différenciation et la reconnaissance des émotions, leur expression et leur gestion inter-individuelle, est selon eux fortement influencé par le contexte social et interactionnel qui lui est associé.

Avec la même idée, Dumas (1948) soutient que les mots « expression émotionnelle » sont « psychologiquement vides de sens si on ne les considère pas dans une vie sociale au sein de laquelle l'expression sera interprétée ». Par exemple le rire est pour lui un fait humain institutionnalisé dont la compréhension nécessite de le replacer « dans son milieu naturel de production qu'est la société ».

Un peu plus tard, dans les années 70 et sous l'impulsion d'Averill (1975), est née la perspective socio-constructiviste pour laquelle la plupart des comportements de l'être humain sont des constructions purement sociales et culturelles (Gergen, 1985).

Pour les constructivistes, la composante sociale est nécessaire pour comprendre l'émotion. Ils considèrent que l'émotion est essentiellement déterminée par des normes, des règles sociales ou des attentes au sein d'un groupe, qui existent dans un environnement social donné. Les émotions seraient les produits d'une culture donnée, construits par cette culture et pour cette culture (Niedenthal, 2007) :

> « [Les émotions seraient] des connaissances acquises par le biais de la socialisation et renforcées au travers des rôles tenus par les individus dans la sociétés. Elles servirait des fonctions sociales et individuelles et ne pourraient être comprises qu'à un niveau d'analyse sociétal. » (*ibid*, cité par Nugier, 2009, p.12)

Rejetant par la même occasion l'idée de configuration spécifique innée et universelle pour chaque émotion, le courant socio-constructiviste s'oppose ainsi à la notion d'émotion de base.

Plus particulièrement, selon Averill (1980), une émotion serait un ensemble de réponses sociales dont la signification serait seulement symbolisée par le label émotionnel que le sujet applique à son comportement. C'est en ce sens que les constructivistes suggèrent que c'est par le biais de pratiques linguistiques que se construiraient les divers états émotionnels.

En effet, même si toutes les langues possèdent des mots pour exprimer les concepts d'émotion, leur nombre varie fortement selon les langues : en face des 550 états émotionnels désignés par des adjectifs anglais (Averill, 1975), d'autres langues, tel le

Chewong (dialecte d'une île de Malaisie) n'en possèdent que 7. Leur catégorisation des émotions ne peut donc être aussi fine, et ce qui est catégorisé comme une seule émotion dans une langue donnée peut correspondre à plusieurs dans une autre langue. Cela a également pour conséquence que certains concepts d'émotion n'ont pas d'équivalent d'une langue à l'autre.

Toutefois, comme le notent Sander & Scherer (2009, p.37) de telles différences culturelles dans la manière de parler des sentiments n'invalident pas nécessairement l'idée que les processus émotionnels soient partagés par les cultures. Pour prouver qu'il n'y a pas ou peu d'universalité, il faudrait en effet montrer qu'il existe de grandes différences interculturelles dans les autres composantes du processus émotionnel (*i.e.* l'*appraisal*, l'expression motrice, les réactions physiologiques, ou encore la tendance à l'action). Or jusqu'à présent, les études de ces autres composantes des émotions tendent plutôt à montrer, au contraire, des variations interculturelles minimes comparées aux différences universelles entre les émotions elles-mêmes (voir Scherer, Wallbott, & Summerfield, 1986 ; Scherer & Wallbott, 1994).

Ainsi, les différentes théories des émotions s'accordent aujourd'hui à dire que les expressions émotionnelles forment un code communicatif dont l'enfant fait l'acquisition par rapport à son expérience, son ressenti. Cette perspective confirme la thèse de certaines théories de l'évolution, selon laquelle les expressions émotionnelles seraient universelles, alors que le code émotionnel, les moments où exprimer telle ou telle émotion, seraient quant à eux acquis.

Les chercheurs ont donc dépassé les idées novatrices de Darwin, au sens où les émotions sont aujourd'hui placées non seulement « comme forme d'adaptation aux problèmes posés par l'environnement immédiat, mais également dans un rôle social, avec une fonction primordiale dans la régulation sociale des comportements » (Nugier, 2009, p.9)

# V. Résumé : Une mise en regard des différentes théories de psychologie des émotions

Dans ce chapitre, nous avons passé en revue les différentes approches générales existantes sur les émotions, et quelques unes de leur modélisations. Leur point de vue diffère selon qu'elles appréhendent les émotions sous l'angle de leurs processus cognitifs, (neuro)physiologiques, de leur fonction adaptative, ou encore de leurs relations au contexte social.

Nous relevons toutefois que la plupart des théories cognitives se fondent sur des principes évolutionnistes et prennent en compte les phénomènes d'activation physiologique. Par conséquent, elles ne se démarquent pas totalement des autres approches des émotions.

De plus, l'approche cognitive peut rendre compte d'un large panel de réactions émotionnelles et de tendances d'action. En effet, contrairement aux théories évolutionnistes, les modèles de l'*appraisal* n'impliquent pas qu'un événement donné soit directement et obligatoirement lié à une réponse émotionnelle précise et prédéfinie (Nugier, 2009). Ces modèles permettent de préciser ce qu'il se passe entre la perception de l'événement et le sentiment subjectif, mais aussi de justifier des variabilités rencontrées en tenant compte de facteurs situationnels et individuels.

C'est pour toutes ces raisons que l'approche cognitive et les modèles de l'*appraisal* sont prédominants aujourd'hui dans le champ théorique de l'étude des émotions.

Quoi qu'il en soit, il ressort des différentes approches de l'émotion une constante, qui consiste plusieurs composantes majeures des émotions, dont les composantes physiologique, cognitive et expressive.

En effet, les émotions semblent nécessairement accompagnées sinon déterminées par une activation physiologique. Ce constat apporte une piste de recherche pour permettre de discriminer les émotions spontanées et les émotions simulées par la mesure de paramètres physiologiques.

Au niveau cognitif, les processus mis en jeu sont déterminants dans le processus émotionnel. De nombreuses études cherchent à établir et comprendre la nature de ces derniers. Scherer notamment a développé un modèle qui détaille précisément leur succession dans le mécanisme émotionnel. Il apparaît également que l'étude de l'expressivité émotionnelle ne peut que gagner à être abordée en relation avec les processus cognitifs qui la sous-tendent.

Enfin, du point de vue de l'expressivité émotionnelle, le rôle communicatif et donc social des émotions semble clairement établi. Ekman a démontré l'universalité de l'expression des émotions mais celle-ci se combinerait à une variabilité culturelle dans des règles d'expressions. Ces règles détermineraient alors si l'émotion peut être ou non exprimée ouvertement. Quant aux approches cognitives, elles apportent, par le biais des évaluations cognitives, une perspective plus adaptée pour décrire l'expression de « mélanges d'émotions » et pour rendre compte des variabilités inter- et intra-individuelles dans les réponses émotionnelles.

Gardant à l'esprit cette variabilité culturelle, mais aussi l'importance des paramètres situationnels et sociaux, nous allons maintenant nous focaliser sur le domaine plus restreint de l'expression des émotions et plus globalement de tout comportement expressif, en particulier sa modalité visuelle à travers les gestes et la face.

# CHAPITRE 2 : LE COMPORTEMENT EXPRESSIF ET SES ENJEUX TECHNOLOGIQUES

L'importance de la partie non verbale de la communication est un fait communément accepté. Il est souvent clamé, au niveau populaire et typiquement sans aucune source scientifique, que la communication non-verbale compterait pour jusqu'à 90% du message. Cette affirmation a sans doute été initiée par les travaux de Mehrabian & Wiener à la fin des années 60 (1967), qui ont pondéré les trois facteurs de la communication que sont le verbal, les facteurs intonatifs (le « paraverbal ») et les expressions faciales, respectivement à 7%, 38% et 55%. 93% de la communication ne serait donc pas du verbal.

En parallèle, les émotions sont exprimées, et c'est bien la raison de leur existence, dans la communication entre humains. Les approches physiologiques ont montré que les changements bio-physiologiques liés aux variations d'états émotionnels sont « récupérés » sous diverses formes, qui vont de la coloration des joues à la voix qui tremble, en passant par la sudation. Par ailleurs, des approches évolutionnistes ont montré que l'expression est aussi le produit d'un contrôle complexe. C'est ainsi que les expressions faciales ont été étudiées, sous l'impulsion d'Ekman en particulier, puis modélisées et même simulées. La pragmatique s'est alors très tôt emparée du problème de la modélisation de la gestualité, en s'ancrant dans une perspective de psychologie cognitive (McNeill, 1992).

Actuellement, l'enjeu du récent domaine de l'« informatique affective » (*Affective Computing*) est de donner un rôle incontournable, lors d'interactions, aux états émotionnels, mais aussi aux intentions, aux affects sociaux, et aux états mentaux. Cette hypothèse est un écho aux propositions de plus en plus étayées des théories psychologiques des émotions, pour lesquelles raisonnements et états émotionnels fonctionnent de manière étroitement liée (*cf.* Chapitre 1). Les interactions considérées peuvent mettre en jeu des humains, bien entendu, mais également des agents virtuels communicants, interagissant avec l'humain, ou entre eux. Modéliser, identifier, voire interpréter le comportement humain en interaction communicative, et doter les agents de comportements similaires, font parties des principaux enjeux technologiques actuels.

Après un bref aperçu de la complexité du comportement expressif et des principales théories de l'expression des émotions, nous pointerons dans ce chapitre les différents enjeux théoriques et technologiques qui interviennent dans un objectif de modélisation du comportement expressif puis nous décrirons la problématique que nous avons adoptée et le paradigme de notre étude.

# I. Le comportement expressif en interaction communicative

> « Même dans les interactions à dominante verbale, le matériel comportemental pertinent se compose de signifiants verbaux bien sûr, mais aussi d'intonations, de rires et de silences, de regards, de gestes, de mimiques, de postures... […] En d'autres termes : la communication est, à l'oral du moins, multicanale. » (Kerbrat-Orecchioni, 1986, p.18)

Par l'organisation même du contenu de son livre « Mind, Hand, Face and Body », Isabella Poggi (2007) met en avant la complexité multicanale de la communication. Ce livre est en effet composé de quatre grande parties, la seconde (« Hands ») et la troisième (« Face »), faisant une analyse détaillée de la communication, respectivement gestuelle et faciale (dont le regard et les sourcils pour cette dernière).

Les canaux de communication les plus largement étudiés et modélisés, sont les gestes, la posture, les expressions faciales et le regard pour la modalité visuelle, et le « verbal » et la prosodie pour la modalité auditive. Toutefois, même si cette dichotomie entre modalité auditive *vs*. visuelle paraît évidente à première vue, elle n'est finalement plus si triviale, à partir du moment où les canaux ne sont plus étudiés séparément. Deux exemples représentatifs sont l'étude des *affect bursts* (Scherer, 1994) que nous développerons dans le chapitre 5 et la notion de « prosodie audio-visuelle » approfondie en particulier par Krahmer & Swerts (2009, entre autres).

Nous reviendrons sur des études marquantes de ces canaux de communication expressive de la modalité visuelle, après avoir expliqué la notion de « prosodie audio-visuelle », puis donné un aperçu de la problématique de la multimodalité, sa terminologie et son contrôle. Nous remarquerons enfin la continuité de l'interaction communicative et la multifonctionnalité du comportement expressif.

## *I.1. Un comportement expressif multimodal*

### I.1.1. La notion de « prosodie audio-visuelle »

Krahmer & Swerts (2009) parlent de « prosodie audio-visuelle ». Selon eux, il est surprenant que l'tétude de la prosodie ne se soit focalisée que sur la modalité auditive, puisqu'une telle perspective unimodale n'est pas complètement représentative d'une situation communicative prototypique, à savoir « un cadre face-à-face dans lequel à la fois le locuteur et l'interlocuteur se voient et s'entendent mutuellement, et portent

attention en continu à des indices auditifs et visuels (*e.g.* Clark & Krych, 2004) »[8]. (Krahmer & Swerts, 2009 p.129).

L'information visuelle, et particulièrement les mouvements des lèvres, sont importants pour la perception de la parole. Cela est illustré par exemple par l'« effet McGurk » (McGurk & MacDonald, 1976). En outre, les chercheurs ont progressivement réalisé que la modalité visuelle de la parole impliquait non seulement les lèvres, mais également le reste du visage. Par exemple Munhall et al. (2004) ont montré que la perception auditive de la parole était améliorée lorsque les mouvement de la tête étaient pris en considération. De plus, le visage d'un locuteur ne contribue pas seulement à la compréhension de la parole, mais sert également « aux fonctions prosodiques traditionnelles comme le découpage et l'emphase » (Krahmer & Swerts, 2009, p.130).

En parallèle, dans le domaine de la synthèse de parole audiovisuelle, les chercheurs se sont intéressés au support visuel de la parole. Bien entendu, l'animation adéquate des mouvements des lèvres est nécessaire (*e.g.* Benoît & Le Goff, 1998 ; Beskow, 1995 ; Massaro & Egan, 1996), mais cela ne suffit pas dans la perspective d'animer une tête parlante qui apparaît relativement naturelle. Des études se sont donc focalisées sur les différentes manières par lesquelles la naturalité de la parole audiovisuelle pouvait être améliorée par des mouvements faciaux appropriés (*e.g.* en montant quel est le mot le plus proéminent (*prominent*) de l'énoncé (*e.g.* Pelachaud, Badler, & Steedman, 1996)). Les résultats de ces études suggèrent que le composant visuel lui-même, qui étaient autrefois associés à la « prosodie verbale », est une valeur ajoutée pour divers aspects de la communication, :

> « Il est ainsi juste de dire que l'information audio-visuelle apparaît importante pour un grand nombre de fonctions communicatives, de telle manière qu'elles peuvent influencer à la fois l'intelligibilité de la parole, et être les signaux d'éléments pragmatiques de haut niveau […]. La prosodie audio-visuelle est donc clairement utile dans les interactions entre humains, et il y a une évidence grandissante qu'elle pourrait rendre les interactions homme-machine plus efficace. »[9] (Krahmer & Swerts, 2009, p.130)

[8] Citation originale : « […] namely a face-to-face setting in which both speaker and addressee see and hear each other, and continuously pay attentionto both auditory and visual cues (*eg.* Clark & Krych, 2004) »

[9] Citation originale : « It is thus fair to say that audiovisual information has been shown to be important for a wide range of communicative functions, as they may influence both speech intelligibility and signal higher-level pragmatic issues [...]. Audiovisual prosody thus serves a clear purpose in human-human interactions, and there is growing evidence that it may make human-machine interactions more effective as well. »

Ainsi, même dans le domaine de la prosodie, particulièrement ancré dans une tradition auditive, la modalité visuelle apparaît être une valeur ajoutée essentielle, à la fois pour l'intelligibilité, et en tant qu'information supplémentaire ou complémentaire.

De plus, des résultats montrent que les éléments à fonction prosodique ont en même temps une fonction affective en signalant des émotions de base (*e.g.* De Gelder & Vroomen, 2000 ; Audibert, 2008 ; ou Barkhuysen, Krahmer, & Swerts, 2010), et des « émotions sociales », comme l'incertitude et la frustration (*e.g.* Barkhuysen, Krahmer, & Swerts, 2005 ; Shochi, Erickson, Rilliard, Aubergé, & Martin, 2008b ; Swerts & Krahmer, 2008 ; Mac, Aubergé, Rilliard, & Castelli, 2010)

De plus, Aubergé & Cathiard (2003) ont montré qu'il était possible d'entendre plus que la simple conséquence acoustique de la modification de la forme du conduit vocal lors de l'expression de l'amusement. Il s'agirait donc plus qu'une intégration audio-visuelle : les expressions émotionnelles seraient contrôlées séparément pour les différentes modalités.

### I.1.2. Précision terminologique sur la multimodalité

Le terme « multimodalité » est couramment employé pour désigner deux phénomènes différents, ce qui implique une ambiguïté terminologique :

- d'une part la multimodalité globale de la communication, c'est-à-dire dans quelle mesure et de quelle manière la gestualité, les expressions faciales, etc., apportent de l'information au langage (*cf.* entre autres : McNeill (1992) et Kendon (2004)) ;
- d'autre part, à un niveau plus local, la multimodalité « intrinsèque », c'est-à-dire la perception de la parole et des autres types d'expression (affectives, etc.) en tant qu'intégration audio-visuelle. Cette multimodalité est typiquement montrée par des tests perceptifs utilisant des stimuli incongruents (McGurk & MacDonald, 1976 ; Massaro & Egan, 1996 ; De Gelder & Vroomen, 2000), c'est-à-dire des conditions de présentation extrayant puis combinant les différentes modalités testées.

Pour éviter toute confusion, Isabella Poggi (2007, p.49 et 50, reprenant Magno Caldognetto & Poggi, 1997) les distingue quant à elle en utilisant respectivement les termes *macrobimodality* (« macro-bimodalité ») et *microbimodality* (« micro-bimodalité »). Alors que la microbimodalité réfère à une même action produite par les deux modalités (visuelle et auditive), la macro-bimodalité concerne des actions indépendantes dans les modalités (*ibid,* p.49, reprenant Messing, 1996) :

> « Par macro-bimodalité (Messing, 1996), je veux parler de la relation existant entre les signaux visuels et acoustiques produits au niveau micro-bimodal, et les signaux

> visuels provenant des autres signaux du corps et produits simultanément : gestes de la main, expressions faciales et mouvements du corps. »[10]

Il est à souligner que ces deux types de multimodalité ont lieu en parallèle, mais à différents niveaux : la macro-bimodalité est planifiée à un plus haut niveau hiérarchique que la micro-bimodalité, puisque « son but général est de transmettre une signification globale, et des parties ou aspects identiques ou différents de cette signification sont distribués à travers les modalités »[11] (*ibid*, p.50).

Nous remarquons ici que la notion de modalité recouvre deux sens (se reporter à Poggi, *ibid,* p.49) :

- la modalité sensorielle, qui dépend des organes sensoriels utilisés par le « récepteur » des signaux de la communication, c'est-à-dire dans notre étude les modalités auditive et visuelle ;
- la modalité motrice, qui dépend des organes du corps qui produisent les éléments communicatifs. Les modalités sont alors, entre autres, et selon la granularité utilisée par les chercheurs : la gestualité, les expressions faciales, le regard, la voix, mais aussi le langage par le biais du conduit vocal.

Dans la suite de ce manuscrit, le sens utilisé pour le terme modalité sera, sans précision supplémentaire, celui de la modalité motrice.

### I.1.3. Multimodalité, répartition de l'information, et problématique du contrôle multimodal

Nous venons d'évoquer une multimodalité globale de la communication. Pour Cosnier & Brossard (1984, p. 5) :

> « Chaque interactant émet (et reçoit) un énoncé total, hétérogène, résultant de la combinaison généralement synergique de plusieurs éléments ».

Mais comment l'information transmise est-elle répartie dans les différentes modalités ? Plusieurs hypothèses sont à considérer :

- cette information est-elle en redondance, en complétion, en cohérence dans les modalités ? (*e.g.* Kendon, 2004 , pour la relation gestualité – modalité verbale)

[10] Citation originale : « By macro-bimodality (Messing, 1996), I mean the relationship among these optical and acoustic signals produced at the micro-bimodal level and the optical signals performed by the other body signals simultaneously produced: hand gestures, facial expression and body movements »

[11] Citation originale : « The general goal is to convey a global meaning, and the same or différent part or aspects of that meaning are distributed across modalities. »

- l'information globale transmise est-elle la somme des informations contenues dans les différentes modalités ? (Ekman & Friesen, 1978)

- ou encore est-elle émergente de ces informations, comme le suggère cette citation de Poggi (2007, abstract) ?

> « Nous parlons par des gestes, des expressions faciales, des regards, des mouvements du corps, des postures, et ces modalités communicatives interagissent les unes avec les autres de manière complexe et subtile. Mais pouvons-nous démêler les différents sons d'une symphonie, les différentes pièces d'une mosaïque ? »[12]

Au niveau des applications technologiques, en reconnaissance comme en synthèse, cette question de l'intégration multimodale apparaît comme un point clé dans l'amélioration des systèmes. Ainsi, Magnenat-Thalmann, HyungSeok, Egges, & Garchery (2005, p.5) ont souligné l'importance de modéliser tous les canaux sensoriels dans les applications IHM, mais également l'importance de l'intégration multisensorielle :

> « Dans un système multimodal […] les canaux de communication sont nombreux : voix, geste, regard, visuel, audio, haptique, etc. Intégrer ces modalités (entrées multimodales) augmente la sensation de présence et de réalisme, et améliore l'interaction homme-machine. »[13]

L'importance de cette intégration est également mise en avant par des sessions de conférence qui lui sont entièrement consacrées, telle la session spéciale « ICME » à IEEE en 2005[14].

Une autre question, encore en suspens, en lien avec celle de la répartition de l'information dans les différentes modalités, concerne le type de contrôle (volontaire ou involontaire - Aubergé, 2002 -), exercé par l'interactant lors de la production des différentes modalités.

Cette problématique du contrôle peut être vue au niveau temporel, c'est-à-dire à quel moment il a lieu par rapport au couplage (ou découplage) de l'information dans les différentes modalités. Le contrôle peut également concerner l'intensité de l'expression, dans une ou plusieurs modalités (selon le moment où il intervient). Le résultat de ce contrôle peut aller de l'inhibition à l'exagération de l'intensité de l'expression.

[12] Citation originale : « We talk by gestures, facial expression, gaze, body movements, posture, and these communicative modalities interact with each other in subtle and complex ways. But can we disentangle the different sounds in a symphony, the different pieces in a mosaic? »

[13] Citation originale : « In a multimodal system […] communication channels are numerous: voice, gesture, gaze, visual, auditory, haptic etc. Integrating these modalities (multimodal inputs) improves the sense of presence and realism and enhances human computer interaction. »

[14] *cf.* http://www.image.ece.ntua.gr/icme2005/ consulté pour la dernière fois le 10/01/11.

Ainsi, trois types de scénario-hypothèse sont possibles :

- seul le choix d'inhiber l'expression de l'information est possible pour l'interactant. Il n'a aucun contrôle sur la répartition de l'information dans les modalités. La redondance est donc totale entre les modalités, *i.e.* elles transmettent les mêmes nature et intensité de l'expression ;

- les modalités transmettent la même valeur d'expression, mais chacune possède un potentiel inhibiteur. Il s'agit d'une redondance modulée : par exemple, de la joie est exprimée globalement, mais cette émotion est transmise à travers la voix et de manière presque imperceptible sur le visage ;

- il existe une possibilité pour l'interactant d'envoyer des informations non redondantes (*e.g.* complémentaires) dans les modalités, et donc de contrôler à la fois la valeur et l'intensité de l'expression pour chaque modalité (*e.g.* exprimer une joie intense par la voix, et en même temps une légère tristesse sur le visage).

Au cours de notre travail, nous nous efforcerons d'apporter des éléments en réponses aux différentes hypothèses concernant la répartition de l'information dans les modalités ainsi que le contrôle des expressions de cette information.

## *I.2. Les différentes modalités visuelles considérées et leur étude*

Un état de l'art des études portant sur les différentes modalités visuelles utilisées en interaction communicative pourrait faire l'objet de plusieurs volumes sans pour autant prétendre à l'exhaustivité, ni éclairer notre problématique, tant ces études ont des approches et des méthodologies variées. Nous partirons donc d'une taxonomie arbitraire[15] de ces modalités visuelles, couramment décrites, concernant l'interaction communicative. Puis, nous nous contenterons de citer pour chaque modalités, à titre d'exemple, des études qui ont influencé les recherches actuelles, et/ou qui nous paraissent particulièrement représentatives. Nous ne nous attarderons pas sur les enjeux théoriques et méthodologiques mis en évidence par ces études, puisque les premiers seront globalement traités dans les parties I.4. et I.3. de ce chapitre, et les seconds seront passés en revue Chapitre 3 II.. Par ailleurs, nous reviendrons sur les modalités réputées « auditives », c'est-à-dire le langage verbal, mais surtout la prosodie, à partir du chapitre 5.

### I.2.1. Les gestes et postures

Les gestes ont la particularité d'avoir été classés de diverses manières, en fonction de l'approche adoptées par la recherche considérée.

[15] En effet, nous considérons ici regard et expression faciale comme deux modalités, en vue du survol des études, mais cette séparation ne fait pas l'unanimité.

C'est dans les années 50, qu'est née la psycholinguistique, et avec elle l'étude du lien existant entre langage et pensée. Le domaine de la gestualité s'est alors intégrée dans ce domaine, en étudiant les gestes soit dans leur relation à la pensée (*e.g.* McNeill, 1992), soit dans leur relation au langage (*e.g.* Kendon, 2004).

> « Les indices observables de la gestualité et de la vocalité sont, ainsi, le reflet de l'adaptation dynamique des mécanismes cognitifs en situation d'interaction. (McNeill, 1992 ; Boyer, 1997) » (Guaïtella et al., 1998, p.13)

McNeill (1992) est parti de l'étude de son domaine, la gestualité, pour d'abord s'attarder sur le lien existant entre gestes et pensées, puis plus particulièrement sur la fonction communicative des gestes et le caractère multimodal de la parole.

A l'exception des gestes autocentrés (gestes de grattage, de confort, etc.) et des gestes ludiques tournés vers les objets (comme jouer avec un stylo), qui se produisent durant la communication mais qui n'ont pas réellement de fonction communicative, les autres gestes sont selon lui en lien avec la parole par leurs significations, leurs fonctions et leurs relations temporelles. Ainsi, au niveau cognitif, ces gestes révèlent l'imagerie de la pensée du locuteur. De là, McNeill (1992) distingue deux types de gestes :

- les *emblèmes,* autonomes et porteurs de sens à eux seuls, qui sont exécutés de manière totalement volontaire par le locuteur et peuvent être utilisés indépendamment de la parole ;

- les *gestes co-verbaux,* qui ne peuvent se substituer à la parole puisqu'ils ne peuvent être interprétés hors contexte, et interviennent spontanément au cours de la parole. Ils entretiennent donc un lien direct avec elle.

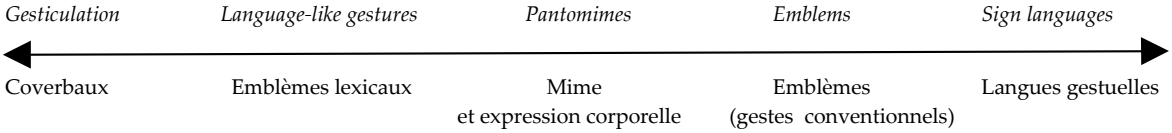
Selon cette classification, c'est la partie des gestes *co-verbaux* qui nous intéresse, ainsi que les remarques concernant leurs processus cognitifs (leur lien avec l'imagerie de la pensée du locuteur, leur caractère automatique, etc.). Par la suite, McNeill (*ibid*) a établi une typologie des gestes selon leur mode de représentation. Les gestes y sont classés dans trois catégories : gestes idéatifs, figuratifs et évocatifs.

Les études de McNeill adoptent une approche relationnelle, au sens où elle part du constat que les relations entre geste et parole sont multiples et complexes. Cette approche s'intéresse particulièrement à la mimo-gestualité, la gestualité déictique, et la gestualité référentielle à travers l'évocation de l'absent (McNeill, 1992, 2005), et l'évocation de l'abstrait (Calbris & Porcher, 1989 ; Calbris, 2003).

L'approche sémiotique des gestes, se préoccupe quant à elle de la manière dont les gestes peuvent avoir une signification d'un point de vue communicatif. Par exemple Kendon (2004, p.104) classe les gestes communicatifs (c'est-à-dire au service de la communication) selon quatre critères :

- leur mode de fonctionnement sémiotique, *i.e.* la manière dont ils font éventuellement référence à un objet ;
- leurs propriétés linguistiques (combinatoire, récursivité, dépendance contextuelle) ;
- leur degré de codification (idiosynchrasique / conventionnel) ;
- leur dépendance à la parole.

Il a ainsi établi un continuum (Kendon, 1988, repris par Mc Neill, 1992, p.37) qui va des gestes nouvellement créés, voire même créés au moment de l'interaction, aux gestes déjà conventionnalisés.




Toutefois, ces types de gestes, tels qu'ils sont décrits par McNeill ou Kendon, ont des fonctions uniquement en lien direct avec le discours et ses objets, et interviennent par conséquent uniquement pendant le tour de parole du locuteur. Or la communication non-verbale ne peut être réduite à ces seules fonctions et aux seuls tours de parole du locuteur.

En effet, Calbris, chercheuse du domaine de la gestualité, explique dans son livre, *L'expression gestuelle de la pensée d'un homme politique* (Calbris, 2003), qu'il existe une « catégorie de gestes non porteurs de "signification" mais gestionnaires de la communication et instruments de sa pragmatique ». Il s'agit pour elle de la « synchronie interactionnelle » qui implique des « phénomènes d'inférences empathiques à base corporelle ». Il serait par conséquent quasi impossible de parler spontanément sans bouger.

Cosnier (Cosnier & Brossard, 1984 ; Cosnier & Bourgain, 1993) définit quant à lui la gestualité coverbale, comme étant conjointe à la parole (mais pas nécessairement en lien direct avec elle), et l'appelle « posturo-mimo-gestualité ». De plus, pour lui, cette gestualité ne se réduit pas aux gestes manuels et céphaliques, mais concerne également les mimiques et expressions faciales, les changements de regards, et les changements de postures.

### I.2.2. Les expressions faciales et mouvements de la tête

C'est en particulier Ekman qui a donné l'impulsion aux études sur les expressions faciales des émotions. Il a d'abord conçu en (1971), avec Friesen et Tomkins, une

première technique dans le but de les décrire : la *FAST (Facial Affect Scoring Technique*[16]). Toutefois, cette technique ne permettait pas de distinguer tous les comportement faciaux visibles. C'est pourquoi à partir de l'examen d'un grand nombre de vidéo, il a développé le *FACS* (*Facial Action Coding System*[17]) avec Friesen et Hager, de manière à identifier les changements physionomiques spécifiques à une contraction musculaire particulière. Il s'agit du système le plus utilisé actuellement pour mesurer et décrire les expressions faciales. Ces dernières sont ainsi analysées comme une combinaison d'*Unités d'Action* élémentaires (*Action Unit = AU*), en référence à l'anatomie musculaire de la face, chacune étant désignée par un chiffre (Ekman & Friesen, 1976).

Ensuite, Ekman a cherché a établir systématiquement des correspondances entre FACS et émotions exprimées. C'est à partir de ces interprétations que notre méthodologie diffère de la sienne (voir Chapitre 3 II.).

Les *AUs* ont été adaptées plus tard aux systèmes automatiques d'animation faciale par leur regroupement en FAP's *(Face Animation Parameters*[18]*).* Les soixante-huit *FAP's* déterminés ont ainsi été catégorisés dans dix groupes, correspondant aux différentes parties du visage, afin de faciliter les traitements. Les points du visage et leurs regroupements, fixés par les *AUs* et les *FAP's*, décrivent la norme MPEG-4 (Pandzic & Forchheimer, 2003).

Par ailleurs, les mouvements de la tête ont parfois fait l'objet spécifique de certaines études, comme celle de Darwin (1872, cité par Poggi, 2007, p. 237), qui chercha à déterminer une origine biologique *vs.* culturelle des signaux « oui » et « non ».

Toutefois, ces derniers sont le plus souvent traités soit avec les postures, soit comme expression faciale.

### I.2.3. Regard et mouvements des sourcils

Le regard, souvent traité parmi les expressions faciales, ou alors couplé avec les mouvements des sourcils, est la première modalité à avoir été étudiée en tant que mode accompagnant la « communication verbale ». Il a été un objet d'étude dans ce cadre dès les années 60, avec les travaux de Goffman (1963). Il constitue en effet une source importante de régulation de la communication (Cosnier & Brossard, 1984; Corraze, 1996). Il est ainsi donné comme tributaire de plusieurs facteurs, notamment sociaux et affectifs, tels l'alternance entre le statut de récepteur et d'émetteur, la

[16] Technique d'Evaluation de l'Affect Facial

[17] Système de Codage de l'Activité Faciale

[18] Paramètres d'Animation Faciale

motivation sociale, les relations entre les interlocuteurs, leur sexe, les traits de personnalité, ou l'état émotionnel.

Dans le domaine de l'*Affective Computing*, Magnenat-Thalmann, HyungSeok, Egges, & Garcher (2005) ont montré que le regard, mais également les mouvements des sourcils, sont des comportements non-verbaux qui ont une influence notable sur la crédibilité des agents communicants animés. Concernant les mouvements des sourcils, leur importance est d'autant plus renforcée en interaction, qu'ils ont été montrés dépendant de l'accentuation de certaines parties de l'énoncé par le locuteur, mais aussi liés à certaines expressions, notamment émotionnelles (entre autres l'incertitude).

## *I.3. Un comportement complexe aux nombreux paramètres*

### I.3.1. Des expressions d'états variés et mélangés

Nous avons vu dans le début de ce chapitre que le comportement expressif était de nature multimodale. En plus de cette caractéristique, la nature même de ce que le comportement exprime est complexe, cela même si seules les expressions émotionnelles sont considérées (*e.g.* Martin, Niewiadomski, Devillers, Buisine, & Pelachaud, 2006).

Comme le remarque Schröder & Cowie (2006), concernant la perception des émotions lors d'études multi-canaux, les taux de reconnaissance des émotions, obtenus avec des données actées peuvent être élevés. Pourtant, les même systèmes rencontrent de grosses difficultés lors de l'utilisation de données naturelles (Cornelius, 1996 ; Russell & Barrett-Feldman, 1999). En parallèle, des études en génération ont très tôt cherché à doter les agents virtuels de comportements émotionnels, en partant d'analyses de visages statiques exprimant des émotions prototypiques (*fullblown*). Bien que le résultat soit reconnaissable, il reste déconcertant par sa non-naturalité (Cowie, 2005). Une des principales raisons est qu'en situation naturelle, les expressions émotionnelles sont le plus souvent de faible intensité et subtiles :

> « Par exemple, les états émotionnels de faible intensité, comme elles apparaissent souvent en dialogue naturel, ne peuvent facilement être décrits par les théories des émotions, qui se focalisent sur les émotions intenses et prototypiques. »[19] (Schröder & Cowie, 2006, p.5)

Cette suggestion est renforcée par une observation de Buisine, Hartmann, Mancini, & Pelachaud (2006). Alors qu'ils cherchaient à évaluer leur système de génération de gestes expressifs pour agents virtuels, il est apparu que la qualité de l'animation était parfois insuffisante pour produire les changements subtiles d'expressivité. La subtilité du comportement expressif est donc actuellement une limite à sa génération.

Par ailleurs, la majorité des états émotionnels exprimés dans les données naturelles sont de nature complexe. Ils sont le plus souvent des mélanges d'états, potentiellement contradictoires, chacun étant éventuellement masqués en partie ou complétement par le locuteur :

> « Les représentations d'émotions fondées sur le lexique ("colère", "tristesse", etc.) ne peuvent capturer facilement les composites et les nuances des émotions, qui sont fréquemment observés dans les données naturelles, *e.g.* quand deux émotions sont

[19] Citation originale : « For example, emotion-related states of low intensity, as they often occur in natural dialogue, cannot easily be described by emotion theories which focus on intense, fullblown emotions. »

> présentes simultanément, ou quand une émotion est exprimée pour masquer celle qui est ressentie »[20] (Schröder & Cowie, 2006, p.5)

Ainsi, Gaver (2009, p.3597) par les termes « brèves, mélangées et masquées » (*fleeting, mixed and masked emotions*), les émotions face auxquelles nous nous trouvons en situation naturelle. Les expressions émotionnelles « naturelles », produites spontanément, diffèrent donc des émotions simulées / actées, qui sont prototypiques et souvent caricaturales.

En outre, Schröder & Cowie (2006, p.4-5) relève que « les nouveaux chercheurs du domaine ont tendance à faire l'erreur de modéliser seulement une courte liste d'émotions de base, sans réaliser que ces émotions ont été proposées dans le contexte spécifique des théories évolutionnistes darwiniennes des émotions »[21]. Selon lui, il peut être utile de cibler les affects étudiés, mais sans utiliser les émotions de base.

D'une manière plus générale, il s'agit même d'aller au delà de l'expression purement émotionnelle, aussi subtile soit-elle. En effet, il est important de modéliser d'autres états affectifs et cognitifs, au moins autant exprimés en interaction communicative que les émotions (*e.g.* De Rosis, 2001, et les études sur les « mind markers », menées en particulier par Poggi -Poggi, 2002 et Poggi, Pelachaud, & Magno Caldognetto, 2004-).

Au niveau des technologies, des auteurs ont par exemple étudié comment la surcharge cognitive ou le doute, pouvaient être reconnus à travers l'analyse du style de langage ou de parole (Berthold & Jameson, 1999 ; Carberry & Schroeder, 2001). D'autres ont montré comment des dialogues simulés pouvaient être enrichis avec d'autres états affectifs, comme l'humeur, le mensonge ou la politesse (Elliott, 1999 ;Ardissono, Boella, & Lesmo, 1999 ; De Rosis, Carofiglio, Grassano, & Castelfranchi, 2003). Le comportement expressif est donc subtil -nous parlerons de micro-expressions-, mais aussi complexe en termes d'états exprimés, à la fois variés et mélangés. Cette complexité apparaît également au niveau des paramètres, en particulier temporels (organisation temporelle, dynamique des mouvements), qu'il est important de modéliser en IHM :

> « Intégrer les émotions dans une IHM demande de considérer les différentes facettes des émotions, avec leurs processus dynamiques, leurs évaluations cognitives et leurs conséquences, leurs expressions comportementales riches et subtiles. » (Pelachaud, 2006, p.442)

[20] Citation originale : « Word-based emotion representations (“anger”, “sadness” etc.) cannot easily capture the composites and shades of emotion that are frequently observed in naturalistic data, *e.g.* when two emotions are simultaneously present, or when one emotion is expressed in order to mask another one that is experienced. »

[21] Citation originale : « Newcomers to the field tend to use short lists of basic emotions, often not realising that these were proposed in the specific context of evolutionary, “Darwinian” emotion theories. »

### I.3.2. L'importance des paramètres temporels

Nous avons vu dans la partie précédente que le comportement exprime le plus souvent un mélange d'états. Cela implique, au niveau temporel, que chacun de ces états peut être déclenché à un moment différent, et évoluer, indépendamment des autres, au cours du temps :

> « L'état affectif d'une personne à n'importe quel moment donné est un mélange d'émotion / attitude / humeur / attitude interpersonnelle, avec souvent un multi-déclenchement d'événements apparaissant à différents moments. Les événements passés et plus récents sont souvent mélangés pour produire un état affectif. […] De plus, les états affectifs sont dynamique et constamment en changement pendant une interaction. »[22] (Campbell et al., 2006, p.xxv)

Cohn (2007) prédit qu'un système ne tenant pas compte de l'information dynamique ne sera pas capable de désambiguïser les expressions avec précision. Pourtant, il remarque que la tendance actuelle des recherches portant sur la reconnaissance automatique des expressions faciales est de se focaliser sur l'occurrence ou la non-occurrence des expressions. Ainsi, la dynamique d'un comportement expressif dépendrait de l'état (ou la combinaison d'états) dans lequel se trouve la personne, et de sa personnalité (Magnenat-Thalmann, HyungSeok, Egges, & Garchery, 2005 ; Cohn, 2007).

Quant à l'organisation temporelle, un des enjeux actuels est de dégager les différentes structures temporelles qui caractérisent les modalités de la communication, et de clarifier leur intégration inter-modale (Schröder & Cowie, 2006). Les paramètres temporels, organisation temporelle et dynamique, de l'ensemble des expressions, sont donc indispensables pour l'étude et la modélisation d'un comportement naturel[23]. Nous reviendrons sur ces paramètres temporels dans la suite de cette thèse. Nous testerons en particulier l'apport de la dynamique en termes d'information sur des icônes gestuelles issues de notre corpus (Chapitre 4), et analyserons différents types d'organisation temporelle des événements vocaux rencontrés dans nos données (Chapitre 6).

[22] Citation originale : « The affective state of a person at any given time is a mixture of emotion/ attitude/ mood/ interpersonal stance with often multi-trigger events occuring at different time. Past and recent evants are often mixed to produce an affective state. […] Furthermore, the affective states are dynamic and constantly in change during an interaction. »

[23] Aubergé (2002a) propose par exemple dans son modèle que, puisqu'il s'agit, selon son hypothèse, d'un contrôle involontaire, la forme prosodique d'une émotion exprimée pendant la parole est superposée temporellement à la parole, sans se contraindre à l'organisation temporelle de la parole. A l'inverse, la forme prosodique d'un affect social (sous contrôle volontaire, *e.g.* la simulation d'une émotion), est quant à elle ancrée, organisée dans et par la structure temporelle du langage. C'est ce qui la distingue de l'émotion involontaire.

## *I.4. Interaction communicative et fonctions du comportement expressif*

La communication est définie différemment selon les chercheurs, en fonction de l'approche adoptée. Toutefois, elle est unanimement décrite comme interactive. Nous partirons de ce constat pour en déterminer les implications, avant de revenir aux comportements expressifs employés au cours de ces interactions. Nous nous interrogerons alors sur les valeurs communicatives possibles de ces comportements, puis sur leurs fonctions.

### I.4.1. Une interaction communicative continue et auto-régulée

Globalement, pour Feyereisen (1994), communiquer signifie exprimer dans un environnement physique, avec des signaux visuels et acoustiques, un état interne. Mais il existe bien d'autres modèles qui cherchent à définir la communication et son rôle, à titre d'exemple :

- Austin (1975) affirme que communiquer est une façon de faire ;
- Poggi (2007) met en avant le caractère multimodal de la communication : nous « parlons » selon elle à travers l'interaction de différentes modalités, et cela avec un but bien déterminé ;
- Grice (1989) suggère que la communication est une « activité intelligente inférentielle » *(intelligent inferential activity*) c'est-à-dire que le locuteur a toujours l'intention d'informer son interlocuteur de quelque-chose, et le lui montre en inférant cette intention.

Malgré ces différences, le fait que les processus communicatifs soient de nature interactive fait l'unanimité :

> « Par interaction [...], on entend à peu près l'influence réciproque que les partenaires exercent sur leurs actions respectives lorsqu'ils sont en présence physique immédiate les uns des autres. » (Goffman, cité par Bachmann, Lindenfeld, & Simonin, 1981, p.127)

Comme le souligne également Kerbrat-Orecchioni (1986), le terme interaction connote l'idée que parler se fait à deux (au moins), et que les deux actants interagissent[24] : le comportement de l'un détermine le comportement de l'autre, et réciproquement.

Pour Poggi (2007, p.40), le processus de communication implique, entre autres, un *sender* (celui qui envoie le message), qui est une condition nécessaire au processus, et un *addressee* (celui à qui s'adresse le message). La particularité de sa théorie est que

[24] Ce sont des « interactants ».

l'*addressee* peut être réel ou fictif. Sa présence n'est nécessaire que dans l'esprit du *sender*.

Nous rejoignons ce point de vue, qui est la raison d'être des agents virtuels expressifs au sein d'Interactions Homme-Machine incarnées. De plus, il permet de justifier la quantité de comportements expressifs produits par les sujets en situation d'IHM (en particulier dans notre corpus, comme nous le verrons) : la machine est ainsi considérée comme un interlocuteur fictif.

Lorsque les humains communiquent en interaction face à face, ils prennent des tours de parole (Duncan, 1972), régulés par le système de *turn-taking*. La principale fonction de ce système est de « séquentialiser en temps réel l'échange d'information entre deux (ou plus) parties communicantes, et de s'assurer que la transmission est efficace » (Thórisson, 2002, p.175)

Lors d'interaction face à face typique, l'information échangée pendant un tour de parole est transmise à travers la parole, les gestes manuels, le langage du corps, le regard, les expressions faciales, et leurs multiples combinaisons (Sacks, 1992 ; McNeill 1992 ; Goodwin, 1981). De plus, Yngve (1970), Sacks, Schegloff, & Jefferson (1974) ou encore Nespolous & Lecours (1986) ont montré que pour mener un dialogue réussi, le processus de *turn-taking* est aussi fondamental que l'information transmise.

> « En théorie, le phénomène de *turn-taking* semble assez facile à définir. Le propos d'une partie entouré par les propos des autres constitue un tour de parole, le *turn-taking* étant le processus à travers lequel la partie tenant le propos à un moment donné est changée. »[25] (Goodwin, 1981, p. 2)

Le *turn-taking* correspond ainsi à un certain nombre de phénomènes de coordination, d'harmonisation, de synchronisation des comportements respectifs des différents interactants, qui permettent la réussite de l'échange communicatif.

Pour permettre ces phénomènes, Duncan (1972, citée par Thórisson, 2002, p.177) propose l'existence d'« indices » de *turn-taking*. Ces indices, variant dans leurs expressions selon les cultures et les individus, seraient générés par chacun des interlocuteurs dans le but d'indiquer aux autres l'état du dialogue, c'est-à-dire si le locuteur veut garder ou céder le tour de parole, et si en parallèle, les interlocuteurs souhaitent prendre le tour de parole ou le laisser.

Par conséquent, pour chaque participant à l'interaction communicative, les comportements expressifs sont continus : ils ne s'arrêtent pas avec la fin de leur tour de parole, loin s'en faut.

[25] Citation originale : « In the abstract, the phenomenon of turn-taking seems quite easy to define. The talk of one party bounded by the talk of others constitutes a turn, with turn-taking being the process through which the party doing the talk of the moment is changed. »

Ces comportements particuliers, produits par les participants d'une interaction, en dehors de leurs tours de parole, ont été décrits pour la première fois par Yngve (1970), sous le nom de « backchannel feedback » (littéralement : « retour d'information de ce qui ce passe derrière »). Le *backchannel* fait ainsi partie des *feedbacks* (« rétroaction ») envoyés par l'interlocuteur. Il peut se manifester sous la forme d'événements vocaux (*cf.* Chapitre 5), de propositions langagières (*e.g.* reformulations, demandes de clarification, etc.), de mouvements de la tête, d'expressions faciales (*e.g.* le sourire), ou encore de regards.

Dans le but d'une modélisation du comportement expressif, par exemple sur un agent virtuel qui interagit avec l'humain, ces *feedbacks,* et plus particulièrement ces indices de *backchannel*, sont fondamentaux pour la crédibilité, la naturalité, et surtout pour le bon déroulement de l'interaction. C'est pourquoi les recherches à ce sujet sont en pleine expansion dans le domaine de l'*Affective Computing* (entre autres, Schröder, Heylen, & Poggi, 2006 ; Heylen, 2007 ; Bevacqua, Heylen, Pelachaud, & Tellier, 2007).

Un des enjeux théoriques sur lequel il est alors nécessaire de se positionner dans un objectif de modélisation, concerne les significations associées à « hors » *vs.* « en tour de parole ». Cette opposition est-elle équivalente au caractère passif *vs.* actif de l'interactant ? Et dans ce dernier cas, « être actif », est-ce le fait d'« être passif », auquel est ajouté quelque-chose, ou est-ce un état opposé à celui d'« être passif » ?

Ainsi, au niveau cognitif, le tour de parole est pour nous dans la continuité du dialogue.

Que nous nous placions en interaction humain-humain ou en IHM, l'information transmise concerne à la fois l'objet de l'interaction communicative et des indices et signaux la régulant. Cela implique une continuité du comportement expressif, en particulier dans les modalités visuelles : l'information transmise ne se limite pas à celle envoyée par les interlocuteurs pendant leurs tours de parole. Comme le souligne Scherer (1994, p.163) concernant les expressions faciales, « la plupart des chercheurs dans ce domaine [des expressions faciales] seraient probablement d'accord sur les fonctions du visage comme un affichage "continu" (voir Buck, 1985) des processus cognitifs et émotionnels en cours. »[26]

### I.4.2. La distinction entre indice et signal au sein de l'interaction communicative

Pour Hauser (1997) il existerait trois modalités de communication animale : les indices (*cues*), les signaux (*signals*) et les signes (*signs*). Nous intéressant aux situations de

[26] Citation originale : « Most researchers in this area would probably agree that the face functions as a *continuous* read-out (see Buck, 1985) of ongoing cognitive and emotional processes. »

communication, nous nous contenterons de développer ici la distinction entre indices et signaux (les signes, au sens de Hauser, n'ayant pas de valeur communicative).

Les indices comme les signaux sont une potentielle source d'information. Toutefois (*ibid*, p.8), alors que les indices sont « actifs » en permanence et ne sont pas sous le contrôle volontaire (*e.g.* la couleur de certains papillons qui signifie « je ne suis pas comestible »), les signaux ne sont pas « actifs » en permanence et peuvent être sous le contrôle volontaire (*e.g.* les cris d'alarme des singes Velvet). Ce critère est cependant complexe à appréhender dès lors qu'il s'agit de caractériser le comportement. Alors qu'est ce qu'un signal en éthologie ?

L'éthologie s'intéresse au comportement dans son évolution. Or, au cours de l'évolution, certains comportements, de l'ordre du stimulus qui engendre une « réponse réflexe » de la part d'un congénère, vont se modifier et s'amplifier. C'est ce qui est appelé la ritualisation. Les deux grandes caractéristiques des actes ritualisés sont l'exagération du comportement originel, et la redondance du contenu du message transmis. En éthologie, un comportement qui a perdu sa valeur, sa signification première, a alors valeur de signal.

Dans cette même idée, pour Seeley (1989, cité par Hauser, 1997, p.10) :

> « Les signaux sont des stimuli qui transmettent de l'information et ont été modelés par sélection naturelle pour cela ; les indices sont des stimuli qui contiennent de l'information mais ne sont pas formés spécifiquement par la sélection naturelle pour transmettre de l'information. »[27]

Ainsi, en interaction communicative, un signal a une signification (arbitraire), par lui-même. C'est un comportement à vertu communicative, qui peut être produit volontairement ou involontairement. À l'inverse, un indice ne va pas avoir de signification par lui-même. C'est l'interprétation, volontaire ou non, que l'interlocuteur en fait, qui lui donne une valeur communicative. De plus, sa présence n'est jamais volontaire.

Dans le domaine de l'*Affective Computing*, Poggi (2007, p.47) définit le signal comme « un stimulus physique lié à une signification dans l'esprit du *sender*, et éventuellement celui de l'*addressee*. »[28] De plus, elle note (*ibid*, p. 48) que des « non-actions » peuvent également avoir valeur de signal : *e.g.* le silence peut être un signal

[27] Citation originale : « Signals are stimuli that convey information and have been molded by natural selection to do so; cues are stimuli that contain information but have not been shaped by natural selection specifically to convey information. »

[28] Citation originale : « The signal is a physical stimulus that in some mind (in the Sender's mind, and, according to the Sender's assumption, also in the addressee's mind) is linked to some meaning. »

puisqu'il peut communiquer du sens (*e.g.* Tannen & Saville-Troike, 1985), de même qu'un visage inexpressif (*blank face*), qui peut parfois être un signal indiquant l'ironie.

L'objectif de notre étude est de comprendre de quelle manière les comportements expressifs rencontrés dans notre corpus transmettent de l'information. La question sous-jacente est ainsi de connaître le statut de ces différentes informations, c'est-à-dire s'il s'agit de signaux dédiés ou d'indices utilisés pour l'expression des états relatifs à l'humain communicant.

### I.4.3. Un comportement communicatif pas seulement émotionnel

Longtemps, et comme dans la tradition des théories des émotions de base, les expressions faciales ont été associées à l'expression émotionnelle. En parallèle, une idée forte des théories sociales des émotions (Chapitre 1 partie IV.), suggérée par Darwin, est que l'expression faciale a essentiellement une fonction communicative, en permettant aux individus d'une même espèce de savoir ce que ressentent leurs congénères. Dans la même lignée, Wallon (1934 ; 1938), puis Malrieu (1952) considèrent les expressions faciales émotionnelles comme une possibilité d'agir sur autrui. Selon l'éthologue Fridlund (1995), les manifestations faciales (pas seulement émotionnelles) sont des signaux qui visent à modifier le comportement du destinataire.

Les expressions faciales sont loin d'être seulement des « purs » signes d'émotion (c'est-à-dire sans aucune fonction sociale) :

> « [Elles ont] à la fois les fonctions duales de régulation communicative et de régulation intra-psychique. Au niveau intra-psychique, les expressions faciales ne sont pas exclusivement des indicateurs de processus émotionnels. Nous trouvons aussi des expressions faciales qui sont des signes de processus cognitifs (un froncement de sourcils indiquant la perplexité), qui peut être ou non des signes d'une réaction émotionnelle (colère) en même temps. »[29] (Kaiser, Wehrle, & Schmidt, 1998, p.86)

Cette citation met en évidence que les expressions faciales peuvent être des signes de processus à la fois émotionnels et cognitifs, sans être incompatible avec une fonction communicative.

> « Les expressions faciales, lors d'interactions, sont très souvent des vecteurs d'informations à l'attention de l'entourage. » (Kaiser, Wehrle, & Schenkel, 2009, p.80)

[29] Citation originale : « the dual functions of communicative and intrapsychic regulation. On the intrapsychic level, facial expressions are not exclusively indicators of emotional processes. We also find facial expressions that are signs of cognitive processes (a frown indicating perplexity), that might or might not be signs of an emotional reaction (anger) at the same time. »

Ainsi, comme nous l'avons vu dans la partie I.4.2. de ce chapitre, chaque comportement peut être un indice ou un signal, selon son potentiel communicatif intrinsèque et sa possibilité d'être produit volontairement ou non.

L'expression faciale, et plus généralement le comportement expressif, auraient donc souvent une fonction communicative et sociale, liée à la présence d'autrui. Mais quel type d'information transmettent-ils ?

Poggi (2007, p.53-60) classe les informations exprimées en trois grandes catégories de significations :

- les informations sur le monde ;
- les informations sur l'identité (biologique et sociale) du *sender* ;
- les informations sur l'esprit du *sender* (« sender's mind »), c'est-à-dire ses croyances, ses buts (y compris ceux liés au *turn-taking*, ainsi que les indices de *backchannel*), et ses émotions.

Plus précisément, nous relevons parmi les potentiels messages transmis, que les expressions faciales et autres comportements (vocaux, gestuels, posturaux, etc.) peuvent être (*cf.* Ekman & Friesen, 1969 ; Kaiser, Wehrle, & Schenkel, 2009, p.80) :

- des signaux régulatifs du discours (*regulator*), *e.g.* les échanges de regards, la réponse d'un auditeur (*backchannel signal*) exprimant par un sourire le fait que le conférencier peut poursuivre sa présentation ;
- des signaux illustratifs, reliés au contenu d'un discours (*illustrator*) (pour donner du poids à son argumentation, ou encore modérer ce qui est dit) ;
- des moyens visant à établir, maintenir, cesser une relation sociale ou à traduire la nature d'une relation ;
- des indicateurs de l'engagement de processus cognitifs (*e.g.* un froncement de sourcils lors d'une réflexion intense) ;
- des indicateurs de l'état émotionnel de la personne (*affect display*).

Les comportements expressifs peuvent donc prendre des valeurs variées, et ainsi avoir différentes fonctions. De ces fonctions découlent les capacités à modéliser et implémenter sur les agents virtuels.

Après avoir approfondi les aspects fonctionnels du comportement, il nous est nécessaire de préciser notre définition du terme « non-verbal ». En effet, notre étude se situe dans le domaine de la communication dite « non verbale ». Cependant, nous fondons l'opposition entre « verbal » et « non-verbal » sur un critère fonctionnel. Cela signifie que pour nous, un comportement non-verbal n'a pas de « fonction de communication construite par le langage » (Loyau, 2007, p. 119). Ainsi, un geste de la main ayant pour fonction de segmenter l'énoncé sera verbal, tandis qu'un

gémissement exprimant un état émotionnel sera non-verbal, et une interjection (pré-lexical) sera à mi-chemin.

## *I.5. Du Feeling of Knowing au Feeling of Thinking*

L'expression *Feeling of Thinking* (*FoT* pour faire court) est un néologisme introduit par Aubergé en 2006 (Loyau & Aubergé, 2006) dont le concept forme le cœur de la thèse de Fanny Loyau (2007), et est au fondement de notre étude. Ce concept dérive de celui de *Feeling of Knowing* (*FoK*), et trouve résonance dans celui de *backchannel*, qui en est un type particulier.

### I.5.1. Qu'est ce que le « Feeling of Knowing » ?

Ce concept, peu développé jusqu'à maintenant, a été décrit pour la première fois par Hart dans les années 60, lors de recherches au sujet de la mémoire et de la méta-mémoire[30]. À l'origine, le *Feeling of Knowing* était défini comme « l'état de croyance dans lequel une information non récupérable sur le moment, mais stockée en mémoire, sera disponible plus tard »[31] (Hart, 1965, cité par Reder & Ritter, 1992, p.435). En somme, le *Feeling of Knowing (FoK*, ou « sentiment de savoir » en français) est le jugement que font les sujets vis-à-vis de leur capacité à reconnaître ou se rappeler des informations non accessibles au moment où ce jugement est fait. Ce phénomène fait penser au « tip-of-the-tongue phenomenon », littéralement phénomène du « mot sur le bout de la langue » dans la langue courante.

Par la suite, le phénomène du *FoK* a été notamment étudié par Shinamura (Shinamura & Squire, 1986), par Reder et collègues à partir de 1987, et en France par Izaute (Izaute, Chambres, & Larochelle, 2002). Il est ressorti de ces études qu'au niveau méta-cognitif, le *FoK* semble faire partie d'un processus général apparaissant automatiquement lorsqu'une question nous est posée. Il déclencherait rapidement un mécanisme de sélection de stratégie de réponse selon la familiarité de la question, puis régulerait le temps de recherche de la réponse (Nhouyvanisvong & Reder, 1998).

[30] « La méta-mémoire concerne les croyances et les connaissances de chacun sur sa mémoire et sur la mémoire humaine (fonctionnement, limitations, capacités, etc.). Elle détermine pour une part la performance de mémoire et intervient dans l'accumulation, la sélection, l'évaluation de l'expérience et dans la construction de l'individualité. » (issu du CV de Marie Izaute, docteur en psychologie cognitive au LaPSCo (Laboratoire de Psychologie Sociale et Cognitive) à Clermont-Ferrand.

[31] Citation originale : « Earlier researchers (Hart, 1965) defined feeling of knowing as only the state of believing that currently unrecallable information will be available (in some form) later because the knowledge is in memory. »

### I.5.2. Prosodie audiovisuelle et « Feeling of Knowing »

Ce sont les néerlandais Marc Swerts et Emiel Krahmer (du laboratoire Communication and Cognition, Tilburg University), qui ont été les premiers, à notre connaissance, à réaliser des études portant sur la relation existant entre *FoK* et prosodie audiovisuelle (Swerts & Krahmer, 2005).

Ils se sont intéressés aux indices audio-visuels de méta-cognition en situation de questions/réponses, en tâche de production comme de perception. Ils ont d'abord porté leur attention sur différents types d'événements qui étaient produits par les sujets auxquels ils posaient des questions factuelles, dans un cadre conversationnel :

- mouvements des sourcils, sourire, nombre de regards vers le bas, le haut, les côtés, « funny face » pour le visuel.

- nombre de mots contenu dans la réponse, délai de réponse, intonation haute à la fin (de type interrogation) et *filled pause* (« hum », « euh », etc.) pour l'audio.

Cette première étude a permis de montrer qu'un certain nombre d'événements visuels et auditifs étaient des indices du *FoK :* diverses correspondances ont pu être établies entre taux de *FoK,* justesse et nature de la réponse, temps de réflexion pré-réponses, et nombre d'indices audio-visuels produits. *À* titre d'exemple, plus le sujet a un *FoK* élevé lors d'une question, moins son temps de réflexion avant de répondre est long, et moins il produit d'indices audio-visuels.

Swerts et son équipe ont ensuite mené une expérimentation portant sur la perception du *FoK* chez les autres, qu'ils ont appelée (*FoAK*, pour *Feeling of Another's Knowing*). Trois conditions ont été testées (vision seule, audio seul et audio-visuel). Il en est ressorti que les observateurs humains sont capables de distinguer une réponse au *FoK* fort d'une réponse au *FoK* faible dans les trois conditions de perception mais plus particulièrement en condition bimodale. D'autre part, les indices visuels, bien que relativement peu fréquents, apparaissent comme très pertinents pour attribuer un *FoAK*.

Les indices audio-visuels des productions de réponses à une question semblent donc être pertinents pour l'interprétation par un observateur humain de « l'indice de confiance » du locuteur envers sa réponse.

### I.5.3. Le *FoT*, généralisation du *FoK* et lien aux comportements de *Feedback*

Nous venons de voir que Swerts et al. ont étudié une situation où il était demandé aux sujets de retrouver une information dans leur mémoire. Comme prévu, le phénomène

de *FoK*, qui correspond aux expressions révélant spécifiquement les processus mnésiques du sujet, est apparu. La principale originalité de cette étude a été de montrer que le sujet exprimait le fait qu'il « sentait qu'il savait » la réponse, de manière multimodale.

Quant à ce que nous observons dans notre corpus, il s'agit du même type d'expressions, mais révélant des processus cognitifs beaucoup plus large que le fait de connaître ou non une réponse. En effet, l'analyse multimodale (voix, parole, langage, expressions faciales, gestualité) de notre corpus a permis de remarquer chez les sujets des expressions non seulement sur leurs états affectifs involontaires ou sociaux (émotions, attitudes, intentions) mais également sur leurs états mentaux (*e.g.* la réflexion à ce que l'on va dire ou la concentration).

C'est pourquoi nous parlerons plus de « sentiment de savoir », mais plutôt de « sentiment de penser » : nous avons ainsi regroupé l'expression de ces différents états sous le terme générique de *Feeling of Thinking* (*FoT*), par analogie terminologique au phénomène du *Feeling of Knowing*.

Mais en quoi consistent les relations qui existent entre notre concept de *FoT* et les comportements de *feedback* et de *backchannel* de l'interlocuteur en interaction communicative ?

Les comportements de *feedback* donnés par l'interlocuteur d'une interaction communicative informent le locuteur sur ses propres états mentaux et affectifs (concentration, recherche en cours d'information connue par le sujet – *i.e. FoK*-, doute, accord, désaccord, inquiétude, satisfaction, etc.). Ils transmettent donc le même type d'informations que les expressions de notre *FoT*, mais sont spécifiquement destinés au locuteur de l'interaction (l'expression du *FoT* ne nécessitant pas de locuteur).

Quant aux comportements à fonction de *backchannel*, ils peuvent être vus comme un type d'expression du *FoT* spécifique à l'interaction communicative et langagière entre deux humains.

# II. Du comportement expressif à son implémentation

Qu'est-il nécessaire de modéliser pour simuler les performances du vendeur qui « sait » vendre, avec des compétences langagières pourtant identiques au vendeur moins efficace ? Tant que la modalité est celle de l'écrit, les affects des agents prennent forme dans leurs stratégies interactionnelles et leur expressivité langagière. Mais ces technologies visent à terme la communication face à face : les clones parlants seront incarnés dans un corps et un visage. Les expressions verbales mais aussi non-verbales sont concernées, avec toute la complexité de la multimodalité faciale, gestuelle et vocale de ces expressions. Cela donne lieu actuellement à de nombreux séminaires et *workshops*, comme par exemple le WACA, qui a lieu tous les deux ans depuis 2006 (dernière édition à Lille en novembre 2010).

Ces dernières années ont ainsi vu l'expansion de la recherche sur le rôle et la fonction des émotions dans les IHM (*cf.* le réseau européen Humaine, et partie II.1.-). En effet, nous avons vu au cours du Chapitre 1, que les émotions sont essentielles pour percevoir, prendre des décisions, interagir avec les autres, etc. Elles jouent un rôle positif dans plusieurs domaines d'application, tels les environnements d'apprentissage, en permettant à l'utilisateur d'avoir une expérience plus agréable et profitable avec le système (Pelachaud, 2006).

Les expressions des états émotionnels, et plus globalement des états affectifs et mentaux, à travers toutes les modalités (partie I. de ce chapitre) sont des vecteurs informationnels de l'interaction. Elles soulèvent un grand nombre de questions théoriques et méthodologiques (entre autres Cohn, 2007), ne serait-ce que par la détermination des variables à observer et à modéliser.

Avant de nous focaliser sur les enjeux de la modélisation du comportement expressif et sur les manière de modéliser et d'implémenter ce dernier, nous allons nous demander en quoi cela est nécessaire, en quoi cela peut être utile, et dans quel but il est éthique de le faire.

## *II.1.Modéliser le comportement expressif : pourquoi ? Dans quel but ?*

### II.1.1. Est-ce nécessaire ?

Du côté des applications, un clone parlant en situation réelle de communication verbale face à face avec un humain résulte déjà d'un casse-tête théorique et technologique. C'est pourquoi il n'est pas inutile de se demander si le doter de la capacité de produire et percevoir les indices émotionnels contenus dans les expressions n'est pas d'une importance périphérique.

Comme le remarque Minsky (1986, cité par Septseault, 2004, diapositive 41), du domaine de l'intelligence artificielle, au MIT :

> « La question n'est pas de savoir si des machines intelligentes peuvent avoir des émotions, mais si des machines peuvent être intelligentes sans avoir des émotions »

Pourtant, lorsque les systèmes artificiels quittent le cadre des situations standard d'interaction personne-machine, traiter les émotions ajoute non seulement de la naturalité, et agit également sur l'efficacité même de la communication.

De plus, lorsqu'un humain se retrouve face à un agent produisant des comportements expressifs, il attribue immanquablement à cet agent une « personnalité » relationnelle à travers ces comportements. Ainsi, des études expérimentales menées par Nass et son équipe depuis 1995, à l'Université de Stanford, ont montré que « les humains répondent aux "personnalités" des ordinateurs de la même manière qu'ils répondraient aux personnalités des humains »[32] (cité par De Rosis, 2001, p.267).

Du point de vue de la génération, cela signifie que le non contrôle des émotions d'un agent virtuel est immanquablement interprété par les sujets humains, pourtant parfaitement avertis de la nature synthétique de l'agent. Ainsi, un couple personnalité / comportement inadapté à la situation, va avoir un effet néfaste sur l'interaction, du entre autres à l'absence de crédibilité de l'agent (Pantic & Rothkrantz, 2003 ; Pelachaud, Peters, Mancini, Bevacqua, & Poggi, 2005).

Sans aller jusqu'à la modélisation inadéquate du couple personnalité / comportement, Schröder & Cowie (2006) donnent deux exemples de « non-fonctionnement » de l'interaction, simplement dus à la modélisation inexistante ou inadaptée d'émotions dans le système :

[32] Citation originale : « people respond to computer personalities in the same way they would respond to human personalities ».

- l'apparition inopinée d'une fenêtre de mise à jour de logiciel (par exemple anti-virus) alors que l'utilisateur est en train de donner une présentation ou effectue une tâche urgente, a plutôt tendance à induire de la panique ou de la colère que de la gratitude ;

- à un niveau plus complexe, un système de dialogue multimodal incapable d'anticiper l'impact émotionnel des informations qu'il donne à l'utilisateur peut être problématique, notamment en situation de relation client (*e.g.* une voix enjouée annonçant à l'utilisateur qu'il n'y a plus de place dans le vol qu'il souhaitait réserver et qu'il ne peut donc pas partir).

En parallèle, Schröder & Cowie (2006, p.194) font remarquer qu'« actuellement, les machines qui interagissent avec des utilisateurs humains ne prennent pas en compte la dimension émotionnelle que les humains s'attendent à trouver en interaction, et c'est une récurrente source de frustration »[33].

Pour toute ces raisons, il est important de modéliser les émotions, au sens large dans les systèmes d'interaction communicative :

> « Les facteurs émotionnels, dans un sens large, sont centraux pour améliorer la naturalité de l'interaction entre les machines et leur utilisateur. »[34] (Schröder & Cowie, 2006, p.194)

« L'informatique orientée émotion » (*emotion-oriented computing*) est l'expression désignant actuellement le domaine qui concerne les technologies qui prennent en considération les émotions (au sens large). À l'origine, il s'agissait du terme *Affective Computing*, « informatique affective ». Ce dernier a été introduit pour la première fois par Rosalind Picard (1997), dans son livre de même titre. Ce livre décrit les premiers travaux sur l'usage des émotions en IHM, et propose des critères de conception pour l'élaboration de systèmes émotionnels. Il met également en avant le fait que puisque l'interaction est de nature bidirectionnelle, son aspect émotionnel peut être vue soit du côté de l'utilisateur (par la détection et la reconnaissance des signes émotionnels qu'il émet), soit du côté du système (simulation des émotions et modélisation des actions possibles du système sur les expériences émotionnelles de l'utilisateur) (Pelachaud, 2006). Par la suite, de nombreux projets de la commission européenne ont contribué à l'extension de l'*Affective Computing* (NECA, ERMIS, SAFIRA, etc.), et notamment plus récemment le réseau européen d'excellence Humaine.

[33] Citation originale : « Currently machines that interact with human users do not take account of the emotional dimension that humans expect to find in interaction, and that is a recurrent source of frustration. »

[34] Citation originale : « Emotional factors in a broad sense are central to improving the naturalness of interaction between machines and their users. »

Le réseau Humaine (HUman-MAchine Interaction Network on Emotions) a été établi afin de préparer, de 2004 à 2007, le terrain scientifique et technologique en vu de créer des systèmes orientés émotion satisfaisants (Schröder & Cowie, 2006).

### II.1.2. Applications technologiques potentielles et éthique

Il existe de nombreuses applications technologiques dans le domaine de l'informatique orientée émotion. Certaines d'entre elles sont déjà développées et commercialisées, alors que d'autres ne sont qu'au stade de développement. Quoi qu'il en soit, toutes sont de formidables moyens de tester les théories et modèles sous-jacents. Nous reviendrons sur cet aspect méthodologique dans le chapitre 3, mais allons donner ici un aperçu de l'étendu de ces applications. Pour plus de détails et d'exemples, nous nous reporterons par exemple à Cowie (2005).

Ces applications orientées émotion concernent :

- les centres d'appel, qui cherchent d'une part à détecter les états affectifs de l'interlocuteur pour adapter la réponse André, Dybkjær, Minker, & Heisterkamp (2004), en donnent plusieurs exemples), d'autre part à générer des voix porteuses d'affectivité ;

- les jeux vidéos (comme Fahrenheit, de la compagnie Gamekult[35] ou Les Sims, de la compagnie Electronic Arts[36]), et les systèmes de narration d'histoires avec affectivité dans la voix, les gestes et les expressions faciales du narrateur virtuel (*e.g.* les applications du projet NECA (Gebhard, Klesen, & Rist, 2004)) ;

- l'éducation et les environnements d'apprentissage, appelés « serious games » par ceux qui utilisent une ergonomie ludique pour l'apprentissage. En effet, en situation pédagogique, montrer de l'« empathie » avec l'état émotionnel de l'apprenant est considéré comme augmentant ses chances de succès (Frasson & Gauthier, 1990 ; Lester, Towns, Callaway, Voerman, & FitzGerald, 2000). Ce constat s'applique ainsi également aux systèmes de *coaching* en vu d'aider l'utilisateur à changer un élément de son style de vie, y compris au niveau de la gestion de ses humeurs ;

- les systèmes cherchant à améliorer la gestion et la production d'émotions chez les individus. Comptent parmi eux les systèmes de communication (tels les synthèses vocales de messages écrits), les systèmes aidant pilotes et conducteurs à gérer leurs états affectifs, et également les applications à visée thérapeutique. Ces applications thérapeutiques peuvent par exemple aider à soigner la maladie de Parkinson[37], et

[35] http://www.gamekult.com/tout/jeux/fiches/J000015364_test.html/, site consulté le 30/09/10.

[36] http://www.sims2.fr/pages.view_frontpage.asp/, site consulté le 30/09/10.

[37] http://www.lsvtglobal.com/ , site consulté le 30/09/10.

d'autres maladie impliquant des troubles émotionnelles et / ou de la communication (comme la dépression ou encore l'autisme). Les synthèses vocales orientées émotion peuvent également permettre aux personnes muettes de « produire » de la parole affective ;

- l'application *marketing* est évidente : il serait alléchant d'un point de vue économique d'être capable de modéliser des agents capables de vendre efficacement. La problématique reste complexe, impliquant de modéliser les techniques *marketing*, c'est-à-dire de trouver les facteurs et paramètres qui entrent en jeu dans la persuasion, l'argumentation, la mise en place d'une relation de confiance, le mensonge, etc. Pour un aperçu plus complet de la problématique, se reporter aux *proceedings* du workshop Humaine « WP8 Emotion in Communication » (Stock & WP8-Members, 2005). Toutefois, il est à noter que ce type d'applications pose un problème éthique.

La dimension éthique de ces systèmes doit en effet être gardée à l'esprit, en particulier pour les applications cherchant à influencer les utilisateurs. Cette dimension a fait l'objet d'un groupe de travail à part entière du réseau Humaine. Un cadre théorique auquel se reporter en a émergé : Goldie, Döring, & al. (2005).

Du côté de l'utilisateur, Gaver (2009) fait remarquer que ces technologies seront acceptées tant qu'elles agiront de manière prévisible. Pour lui, la réticence à ces technologies réside dans la peur qu'ont certains utilisateurs, dans la capacité qu'a (ou qu'aura peut-être dans le futur), la machine à contrôler certains aspects de leur vie et / ou à les influencer.

### II.1.3. Quel est le but ?

La réponse à cette question diverge selon les personnes. Pour certains, elle peut être d'imiter les capacités de l'humain, soit uniquement en termes de résultats, soit en termes de résultats mais également en termes de manière d'y arriver, c'est-à-dire au niveau des processus. Pour d'autres, l'objectif est de faire mieux que l'humain.

Dans les deux premiers cas, le contexte de l'IHM est un moyen d'identifier et / ou de tester les paramètres acoustiques et visuels qui transmettent attitudes et émotions (*e.g.* Cowie, 2005 ; 2009). Les technologies servent ainsi à répondre aux questions théoriques.

Pour le réseau Humaine (Cowie, 2005), l'objectif est de faire aussi bien que l'humain, c'est-à-dire de chercher à produire des IHM aussi riches que les interactions entre humains.

Mais en reconnaissance, est-il nécessaire de faire en sorte que le système fasse les mêmes « erreurs » que l'humain (*e.g.* percevoir les mêmes illusions d'optique, ou dans le domaine émotionnel, ne pas percevoir les émotions masquées par un humain au visage impassible), ou alors doit-il être aussi précis que possible, voire même doit-il, PEUT-il être plus compétent que l'humain ? Pour Schröder & Cowie (2006), la réponse à cette question dépend du domaine d'application.

Au niveau de la génération, l'objectif est une modélisation plus écologique, et la question se pose en terme de crédibilité (*believability*) de l'environnement virtuel, de l'ACA, et plus précisément de son comportement expressif. Pour Magnenat-Thalmann, HyungSeok, Egges, & Garchery (2005), un comportement est crédible par le réalisme de sa génération, mais également par la pertinence et la cohérence des émotions, de la personnalité, de l'intention, etc., que le concepteur cherche à faire attribuer à l'agent, et leur adéquation à la situation.

Pour eux, malgré cette volonté d'imiter l'humain, à commencer par l'apparence anthropomorphique des agents, les résultats actuels de la modélisation des émotions et de la personnalité sur des ACAs ne sont pas très convaincants jusqu'à maintenant. Ils l'expliquent par différentes raisons :

- les modèles psychologiques de l'émotion et de la personnalité ne sont pas finalisés ;

- lorsque l'objectif est la simulation informatique des émotions ou de la personnalité, la tendance est à utiliser le modèle le plus adapté à une simulation computationnelle, même si ce modèle n'est pas nécessairement la meilleure représentation des émotions/de la personnalité. La plausibilité des processus simulés est ainsi mise de côté, et seule l'obtention d'un résultat compte alors ;

- même s'il existait un modèle parfait des émotions et de la personnalité, d'autres enjeux, tels que l'apparence de l'agent, ses capacités, les éléments environnementaux, etc., interfèreraient avec notre impression sur l'état émotionnel de cet agent.

Ainsi, dans le but d'imiter les compétences de l'humain, y compris, dans la mesure du possible, les processus cognitifs sous-jacents, l'effort doit avant tout être porté sur la conception du modèle à travers la recherche fondamentale. Le domaine étant récent, une attention particulière doit être portée sur les différents paramètres à modéliser. C'est seulement si le modèle se fonde sur ce type de descriptions, que son implémentation peut prétendre être crédible en génération, et efficace en reconnaissance.

### II.1.4. Vers un objectif de crédibilité

Pour Magnenat-Thalmann, HyungSeok, Egges, & Garchery (2005), les paramètres à prendre en compte doivent permettre la crédibilité de l'interaction. Le niveau de crédibilité de l'interaction est à la fois dépendant de l'immersion perceptuelle que le système permet, et de l'adéquation de l'interaction à ses propres buts, et donc globalement à la situation:

> « Nous croyons que l'immersion perceptuelle est induite par l'intervention d'intentions, d'émotions et de la personnalité, toutes orientées par des buts. »[38] (*ibid*, p.3)

Ainsi, selon eux, trois éléments entrent en jeu pour obtenir un environnement ou un agent virtuel crédible : l'immersion (spécifique à l'environnement virtuel), la « présentation » et « l'interaction ». La « présentation » concerne le réalisme du monde ou de l'agent virtuel, qui passe par exemple par le fait d'ajouter une modélisation de la texture de la peau à une tête parlante (Courgeon, Martin, & Jacquemin, 2008). « L'interactivité » est quant à elle un des enjeux les plus importants de l'environnement virtuel, puisque les personnages virtuels doivent réagir de manière à ce que leurs buts, désirs et attitudes soient compris, qu'ils soient montrés ou « réels ».

En somme, pour qu'une animation faciale soit crédible, les paramètres les plus importants sont l'apparence, la pertinence du comportement, à la fois dans son adéquation au contexte et au *background* culturel, et au niveau de l'expressivité des mouvements (à travers leur amplitude, leur tempo, etc.), et la synchronisation entre le verbal et les comportements non-verbaux (Magnenat-Thalmann, HyungSeok, Egges, & Garchery, 2005).

Rappelons également (*cf.* partie I.4.1.) que lors d'une interaction avec un agent virtuel ou une tête parlante, le dialogue lui-même est seulement une petite partie, d'où l'intérêt d'étudier le comportement expressif des interactants qui a lieu entre les séquences de parole.

## *II.2. La corporéisation du comportement expressif*

### II.2.1. Les agents virtuels communicatifs

C'est en dotant de *Belief, Desire & Intention* (BDI -*cf.* Chapitre 1) des avatars, des *chatterbots*, en interaction communicative avec l'humain, que sont nés les *Affective Agents* de Picard (1997) au Medialab-MIT, les *Embodied Conversational Interface Agents*

[38] Citation originale : « We believe that the perceptual immersion is invoked by goal-oriented intervention of intents, emotions, and personality. »

de Cassell, Sullivan, Prevost, & Churchill (2000) ou encore les *Believable Social and Emotional Agents* du projet Oz (Reilly, 1996).

Les ACAs offrent aujourd'hui un nouveau champ d'application aux simulations d'expressions faciales et de comportements expressifs. Décomposons l'expression Agent Conversationnel Animé pour déterminer la nature de ce type de technologie (Sansonnet, 2006, diapositive 4) :

- un *agent* est un « composant logiciel capable de raisonnement sur les représentations qu'il a de l'état courant du système, du profil de l'usager et de la tâche à accomplir » ;
- *conversationnel* signifie que le composant logiciel est « capable d'interactions multimodales avec l'usager » ;
- *animé* implique que le composant logiciel est « doté d'une apparence effective face à l'usager : personnification / *embodiement* ».

> « Les Agents Conversationnels Animés sont un type d'interface visant à transférer les propriétés de l'interaction homme-homme, avec toute leur richesse, à l'interaction homme-machine. Il s'agit de personnages virtuels multimodaux capables de communiquer avec l'utilisateur par de multiples modalités : la voix, les expressions faciales, la direction du regard, les gestes des mains, les postures corporelles, etc. » (Buisine et al., 2006, p. 622).

L'expression ACA correspond au terme anglo-saxon ECA, pour « Embodied Conversational Agent », qui aurait pu être traduit littéralement « Agent Conversationnel Incarné », c'est-à-dire ayant une « apparence ». La personnification, la corporéisation (*embodiement* en anglais) sont des éléments communs à ces agents virtuels. Leur degré peut cependant être variable.

Alors que certains agents virtuels sont à peine humanoïdes, comme l'agent de l'aide de Microsoft Office, qui a par défaut est un trombone, d'autres sont des têtes (voire même des corps complets) anthropomorphiques simulant des comportement qui se veulent aux plus proches de ceux de l'humain (*e.g.* GRETA - Poggi, Pelachaud, De Rosis, Carofiglio, & De Carolis, 2005 -, ou MARC - Courgeon, Martin, & Jacquemin, 2008 -).

Selon Sansonnet (2006) les Agents Conversationnels Animés peuvent avoir trois rôles principaux :

- « assistants » : ils sont présents pour aider et guider les utilisateurs à la compréhension et à l'utilisation d'applications et de services informatiques (*e.g.* au laboratoire LIMSI, le projet DAFT a pour objectif de créer des modèles d'agents assistants) ;

- « partenaires » (souvent appelés « avatars ») : ils sont des personnages virtuels dans des environnements virtuels (*e.g.* la communauté tridimensionnelle virtuelle « Second Life ») ;

- « tuteurs » : ils aident l'apprentissage dans les Environnements Interactifs d'Apprentissage Humain (EIAH).

L'utilisation d'agents virtuels avec ces rôles divers est actuellement en plein essor, cela dans de nombreux secteurs (cinéma, jeux-vidéo, assistance diverse, apprentissage, formation, etc. -*cf.* II.1.2.-). Dans ce domaine, l'objectif est la plupart du temps de modéliser un comportement écologique, et non prototypique : les chercheurs essayent de doter les agents de comportements expressifs selon une situation, mais aussi une personnalité particulières, qui peuvent être propres à une culture donnée (Poggi, Pelachaud, De Rosis, Carofiglio, & De Carolis, 2005).

### II.2.2. Un exemple d'agent virtuel : GRETA

GRETA est un exemple d'ACA anthropomorphique parmi les plus aboutis (voir Figure 6 - (Pelachaud, 2010) -). Créé au LINC (Paris) par l'équipe de Pelachaud, il y a une dizaine d'année, l'agent virtuel GRETA a été particulièrement développé ces dernières années grâce à de nombreuses collaborations, en particulier au sein du réseau européen d'excellence Humaine (*cf.* II.1.).

Figure 6: L'ACA GRETA

Cet agent a les capacités de communiquer en utilisant non seulement la parole mais aussi les comportements non-verbaux. Selon leurs principes, toute fonction communicative est composée de deux entités : un signal (action musculaire) et un sens (ensemble de buts, de croyances que l'orateur veut transmettre à l'interlocuteur). Ce

signal peut être une expression faciale, un regard, un mouvement de tête, etc. Cependant, le sens correspond à la valeur communicative d'un signal.

Pelachaud et collègues ont également développé un deuxième système, qui prend en entrée un texte contenant des informations, suivant la définition du langage de représentation APML, sur les fonctions communicatives qui accompagnent le texte. Le système interprète alors le texte en instanciant les fonctions communicatives avec leurs expressions faciales respectives. L'*output* ainsi est fait de deux fichiers, l'un contenant l'audio et l'autre les paramètres pour l'animation faciale.

Depuis quelques années, les efforts sont particulièrement destinés à varier et rendre pertinents les comportements expressifs produits par GRETA. Il s'agit également de la doter d'une personnalité et d'un style de communication particuliers. (Poggi, Pelachaud, De Rosis, Carofiglio, & De Carolis, 2005).

### II.2.3. De la modélisation à l'implémentation

Sansonnet (2006) distingue deux grandes problématiques scientifiques (qui peuvent être mélangées dans certaines approches) :

- les « modèles d'humains », dont l'objectif est la pertinence écologique de la modélisation (de la physiologie, des expressions ou des comportements des humains) ;
- les « modèles d'agents », dont le focus est mis sur les performances interactionnelles des modèles. Parmi eux se trouvent en particulier les modèles d'interaction avec les usagers (*e.g. l*e dialogue homme/machine), et les modèles d'agents cognitifs (*e.g.* la logique BDI et le modèle OCC)[39].

Les modèles d'humains impliquent un type de techniques particulières pour l'implémentation, cherchant à reproduire avec fidélité les comportements observés. Par exemple :

- la capture de mouvements ou l'extraction de suivi, comme le modèle à billes utilisé par le laboratoire GIPSA-lab (voir Figure 7, Raidt, 2008, p.86). Elle permet de reproduire avec exactitude des mouvements réellement produits par un sujet, enregistrés en trois dimensions ;
- l'analyse-resynthèse (*e.g.* Martin et al., 2006). Le principe consiste à d'abord annoter précisément, souvent manuellement et à différents niveaux de description (les signaux multimodaux envoyés, mais aussi leur fonction communicative, les émotions

[39] Quelques exemples : le *Chatterbot* : A.L.I.C.E. (Richard S. Wallace / www.alicebot.org) ; le site de renseignement médical OddCast ; les agents éducatifs de « La Cantoche » (*e.g.* l'agent Laura sur le site d'EDF ; les WebLea bots (LIMSI, Projet Nice).

exprimées, ou encore le contexte) un corpus vidéo de données réelles ; puis d'utiliser ces différents niveaux d'annotation pour animer un Agent Conversationnel Animé.

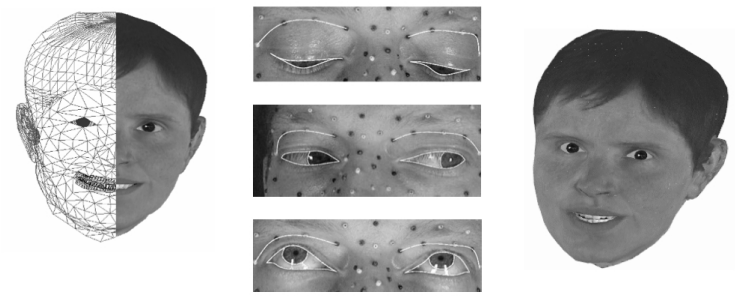


Figure 7: La tête parlante du GIPSA-lab, et son modèle à billes

Quelle que soit la problématique adoptée, il est nécessaire de modéliser à la fois la perception, la génération et l'interaction, pour qu'une interaction entre un ACA et un humain soit perçue comme naturelle par ce dernier (Schröder & Cowie, 2006). Au niveau de l'interaction, il s'agit de modéliser un agent locuteur, mais surtout interlocuteur (*cf.* l'importance de ce qui se passe hors tour de parole, partie I.4.). Il doit être capable de produire des comportements de *backchannel* (*e.g.* pour exprimer s'il a compris ou non les propos de l'utilisateur). Le comportement expressif de l'agent virtuel doit être généré dans les différentes modalités (voix, gestes, expressions faciales, etc.) de manière riche et avec un contrôle précis. De plus, une attention particulière doit être portée sur l'intégration multimodale. Enfin, l'ACA doit être capable de « percevoir » l'utilisateur, les événements, ou les autres agents, afin de produire un comportement adapté à la situation.

Cette situation (au sens large) peut évoluer au cours de l'interaction communicative. Cohn (2007), propose pour s'y adapter, de copier l'approche itérative observée chez l'humain en interaction. En effet, nous faisons en continu des inférences à propos des émotions, des états mentaux, des croyances, des intentions, etc. de nos interlocuteurs, à partir de leurs comportements et des événements extérieurs que nous percevons. L'agent virtuel devrait pouvoir « percevoir automatiquement et modéliser dynamiquement un large spectre de comportements multimodaux provenant de plusieurs personnes, d'évaluer le contexte, de développer des représentations des différences individuelles, et de formuler et tester des hypothèses à travers l'échange de signaux communicatifs. » [40] (*ibid*, p.13)

[40] Citation originale : « automatically detect and dynamically model a wide range of multimodal behavior from multiple persons, assess context, develop representations of individual differences, and formulate and test tentative hypotheses though the exchange of communicative signals."

## *II.3.La modélisation de la variation au sein d'une communication située*

Dans les premières parties de ce chapitre, nous avons entrevu la variété des paramètres situationnels d'une interaction communicative, ainsi que l'influence que ces paramètres peuvent avoir sur l'interaction. Nous allons revenir dans cette partie sur ces différents paramètres et sur la manière dont ils peuvent influencer le comportement expressif. Puis, nous porterons notre attention sur les variations inter-individuelles et la démarche que nous adoptons pour les modéliser.

### II.3.1. Les paramètres d'une communication située

Pour Poggi (2007), la différence entre une communication et une communication réussie réside dans la prise en compte efficace du contexte. Elle a donc établi un modèle du contexte dont le tableau récapitulatif (voir Tableau 1) en indique les différents paramètres à prendre en compte.

Tableau 1: Traduction du modèle du contexte de Poggi (2007, p.83)

<table>
<tr><td>Que communiquer?</td><td colspan="6">But communicatif: quoi communiquer à qui</td></tr>
<tr><td rowspan="15">Comment communiquer ?</td><td rowspan="15">Les conditions du monde: quel est le contexte</td><td rowspan="2">Capacités internes</td><td colspan="4">Compétence linguistique</td></tr>
<tr><td>Conditions pathologiques</td><td colspan="3">transitoires</td></tr>
<tr><td rowspan="13">Conditions externes</td><td colspan="4">permanentes</td></tr>
<tr><td rowspan="3">Contraintes physiques</td><td colspan="3">identiques / temps différent</td></tr>
<tr><td colspan="3">identiques / espace différent</td></tr>
<tr><td colspan="3">Médium possible</td></tr>
<tr><td rowspan="2">Type de rencontre</td><td colspan="3">Rencontre en vue d'un service</td></tr>
<tr><td colspan="3">Echange personnel</td></tr>
<tr><td rowspan="2">situation sociale</td><td colspan="3">publique / privée</td></tr>
<tr><td colspan="3">formelle / informelle</td></tr>
<tr><td rowspan="9">Modèle de l'interlocuteur</td><td rowspan="3">Paramètres cognitifs</td><td>Compétences linguistiques</td></tr>
<tr><td>Base de connaissances</td></tr>
<tr><td>Capacité d'inférence</td></tr>
<tr><td rowspan="2">Paramètres de personnalité</td><td>Buts caractéristiques</td></tr>
<tr><td>Emotions caractéristiques</td></tr>
<tr><td rowspan="3">Relation sociale</td><td>Force de la relation</td></tr>
<tr><td>rôle dans la relation</td></tr>
<tr><td>Affects</td></tr>
</table>

Ce modèle s'intéresse au contexte dans un sens très large. Il pose la question « quels paramètres peuvent influencer ce que le *sender* va communiquer, ou comment il va le communiquer ? ». Selon Poggi, outre le but communicatif du *sender*, c'est-à-dire « que communiquer ? et à qui ? », le contexte comprend les capacités internes du *sender*, et les conditions externes. Nous allons seulement approfondir ces dernières dans cette partie, même si nous considérons comme elle que les deux premiers paramètres font partie du contexte. Pour notre part, nous nous interrogerons ici sur ce qui peut

influencer le comportement expressif, et surtout sur les paramètres situationnels qu'il est nécessaire de modéliser dans un système d'IHM.

Pour Borod et al. (2000), les caractéristiques du sujet, et plus globalement les facteurs situationnels peuvent influencer le « choix » de transmettre de l'information dans un canal de communication en particulier, alors saillant. Nous pensons que plus que ce choix du canal de communication, ces paramètres vont influencer globalement ce qui est exprimé et comment.

De nombreux chercheurs s'intéressent aux éléments du contexte qui influencent l'expression émotionnelle (De Rosis, 2001) et son déclenchement (Ortony, Clore, & Collins, 1988 ; Elliott, 1998 ; Poggi, Pelachaud, & De Carolis, 2001).

Selon De Rosis (2001), l'environnement externe (*e.g.* le lieu, le type de situation -IHM ou interaction entre humains-) stimule ou non l'émotion. Plus particulièrement, l'environnement social (*e.g.* la personnalité et / ou l'état affectif des interlocuteurs, leur rôle social, la relation établie avec eux, mais aussi la culture et le groupe social d'appartenance du *sender*) influence la régulation de l'émotion.

Nous retrouvons les mêmes types d'éléments que dans le modèle de Poggi (ci-dessus), qui porte pourtant sur la communication en général. Ainsi, au-delà de l'expression des émotions, ces même éléments du contexte peuvent influencer plus largement l'expression du *FoT*. Par exemple un présentateur du journal télévisé n'exprime pas de la même façon ses états mentaux et affectifs (du *FoT*) selon la situation (pendant le journal ou dans sa famille). Suchman parlait déjà, en 1987 (cité par Gaver, 2009), d'« émotions situées ». Quant à nous, nous parlerons plus généralement de « communication située ».

Notons que la situation de la communication englobe le temps. En effet, l'état émotionnel d'un agent à un instant donné dépend de son état mental à ce moment, des événements qui se sont passés immédiatement avant, et aussi de son état émotionnel antérieur (De Rosis, 2001). Les émotions sont donc temporellement situées. Ces paramètres font partie du contexte cognitif.

Cela implique qu'il est important de s'intéresser au fonctionnement cognitif et aux causes de l'expression « externe », d'autant plus que les émotions peuvent être déclenchées de manière endogène, et donc pas nécessairement par un événement externe (*ibid*).

Dans les technologies d'interaction face à face, réagir de manière appropriée à la manifestation d'un état affectif de l'interlocuteur, nécessite non seulement de le reconnaître, mais également de raisonner sur les causes de cet état. Cette réflexion n'est pas triviale : De Rosis (*ibid*) note que l'interprétation de la signification subtile

d'énoncés n'est ni immédiate, ni unique. Par exemple Ekman & Friesen (1978) ont montré qu'un même signal pouvait se produire dans des situations très différentes.

Une solution possible à ce problème d'ambiguïté est d'essayer de trouver une interprétation plausible en combinant toutes les informations disponibles : combiner signaux verbaux et non-verbaux, et raisonner à la fois sur l'événement, l'action ou l'objet qui a probablement produit l'état perçu, et sur le mécanisme qui a déclenché cet état et la décision de le manifester (ou de partiellement le cacher). Pour cela, il est nécessaire que les recherches s'intéressent à l'arrière-plan cognitif des comportements expressifs.

Revenons à l'environnement social. Son influence directe réside dans le fait que la simple présence d'une personne influence le comportement. L'influence sera ensuite différente selon l'identité de la personne et sa relation sociale avec le *sender.* C'est ce qui est souvent appelé « régulation interpersonnelle » (*e.g.* Schröder & Cowie, 2006 ; Cohn, 2007). Parmi les phénomènes de régulation interpersonnelle se trouvent (Cohn, 2007) :

- la synchronisation, ou cohérence, qui réfère au fait que des individus en interaction vont bouger simultanément, et avec des actions faciales, des valences affectives, etc. similaires ;
- la réciprocité, qui correspond au fait que le comportement d'un des interlocuteurs va être fonction de celui de l'autre interlocuteur ;
- le *Coordinated Interpersonal Timing* (CIT, littéralement « organisation temporelle interpersonnelle coordonnée »), qui réfère à l'ajustement de l'organisation temporelle liée au *turn-taking* (pauses, etc.) entre participants d'une interaction sociale.

Au niveau des technologies, Cowie (2005) relève qu'un agent ne peut paraître « engagé émotionnellement » dans l'interaction, que s'il est doté d'un genre d'empathie. Cela signifie qu'il doit être capable de « comprendre » quel état émotionnel peut encourager une personne, et comment les états (affectifs et / ou cognitifs) de cette personne peuvent être modifiés par ses actions.

De plus, l'environnement social influence également l'expression par le biais de la culture et du groupe social d'appartenance du *sender*.

La culture est un ensemble de croyances partagées, d'attitudes et de comportements qui sont transmis d'une génération à l'autre (Barnouw, 1985). Quant au groupe social, (Hess & Kirouac, 2003, p.368) le définissent comme « les membres d'une catégorie sociale qui partagent une caractéristique commune associée avec des croyances ou des

rôles partagés concernant l'affectivité »[41]. Le groupe social est ainsi le plus souvent, un sous-groupe de la culture d'appartenance.

L'appartenance à un groupe social peut influencer le processus d'évaluations, tandis que le groupe lui-même, lors de l'interaction, peut influencer le comportement expressif par le biais des *display rules* (*e.g.* Matsumoto, 2007) : nous pouvons adopter une émotion plutôt qu'une autre afin de nous conformer aux normes d'expression des émotions, variables selon les groupes.

Cette notion de *display rules* (littéralement « règles d'exposition ») a été introduite par Ekman & Friesen (1971). Elle correspond à des règles d'expression implicites, qui peuvent varier selon la culture et les groupes sociaux. Ces règles prescrivent quand et comment montrer les émotions et attitudes dans les différents contextes. Il s'agit donc d'un contrôle volontaire (comme la simulation ou l'inhibition -*cf.* partie suivante-) des expressions.

### II.3.2. Le contrôle volontaire ou involontaire des expressions multimodales en interaction communicative

Par l'existence même des *display rules*, Ekman pré-suppose que ce que nous montrons est largement contrôlé. Kaiser, Wehrle, & Schenkel (2009, p.82-83) relève au moins quatre règles, liées chacune à un type de contrôle :

« - modérer l'intensité de ce que l'on montre [...] ;
- intensifier, au contraire, ce que l'on exprime [...] ;
- neutraliser ce qui est ressenti et ce qui est montré [...] ;
- masquer l'affect ressenti en montrant un état différent de celui du moment […]. »

Nous noterons que ces quatre types de contrôle peuvent être ramenés à deux possibilités d'action sur les expressions : la simulation ou l'inhibition.

Le contrôle volontaire et involontaire de l'expressivité a été modélisé par Aubergé (2002b). Pour elle, la parole est une modalité incontournable de l'expression des affects humains. Selon son modèle (*ibid*), différents processus véhiculent les informations relatives aux différents niveaux d'affects. Elle distingue les émotions des attitudes (ou affects sociaux) et de l'expressivité (*i.e.* les stratégies linguistiques). L'hypothèse principale est que dans une communication face-à-face, différents niveaux cognitifs d'informations affectives sont intégrés dans les différentes modalités qui entrent en jeu en interaction communicative (voir une schématisation figure 8) :

[41] Citation originale : « members of a social category who share a common characteristic associated with shared beliefs or roles regarding emotionality. »

- Les émotions sont déclenchées par un contrôle involontaire (sauf dans le cas de la simulation) et sont exprimées dans et par la voix et les modalités d'ordre visuel. Leur processus d'expression, vraisemblablement inné, est seulement soumis à l'apprentissage à travers le contrôle de l'intensité de l'expression (de l'inhibition à l'exagération). Les expressions des émotions sont régies dans un temps cognitif des événements émotionnels, organisé par les événements de causalité des variations émotionnelles induisant l'expression, internes ou externes au contexte communicationnel.

- Les attitudes, affects sociaux intentionnels, sont déclenchées par un contrôle volontaire et sont exprimées directement, en particulier par la prosodie et les expressions faciales. Cette dernière est inscrite dans le temps cognitif du langage. Les concepts et les morphologies prosodiques sont acquis, en fonction de la langue et de la culture. Ce type d'affects est celui qui occupe le plus largement le canal de la parole (*e.g.* Campbell, 2004 ; Aubergé, 2002b ; Shochi, Erickson, Rilliard, Aubergé, & Martin, 2008 ; Wichmann, 2002) et permet au locuteur de fournir des informations quant à ses intentions (doute, évidence, politesse, autorité, confiance, etc.).

- Le niveau le plus complexe des affects est celui de l'expressivité langagière, résultant de méta-processus sur l'ensemble des structures linguistiques (prosodie, lexique, morphologie et syntaxe).

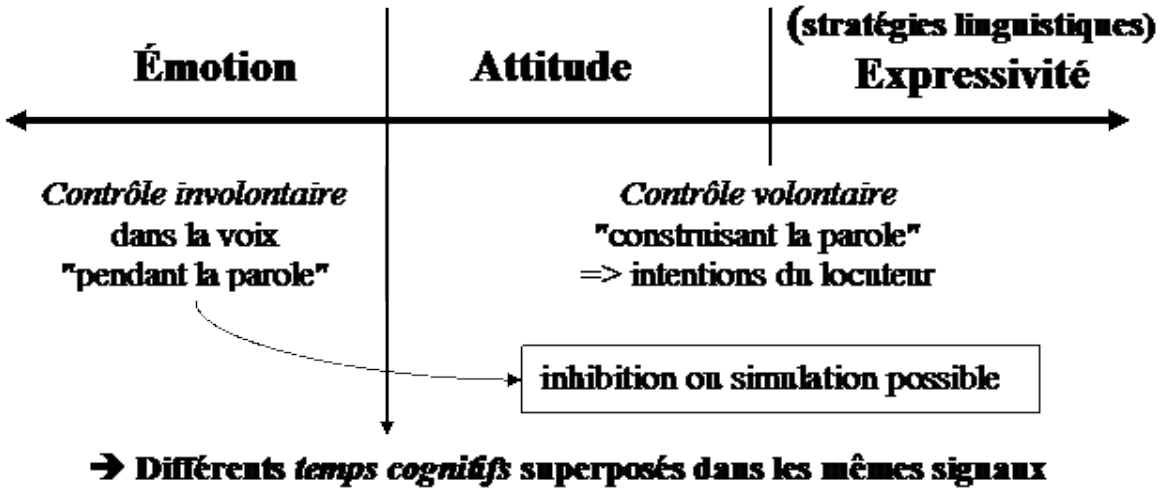



Figure 8: Les différents indices/expressions pour la communication (Aubergé et al., 1998-2011)

C'est dans ce cadre théorique, et donc dans cette distinction entre émotions sous contrôle involontaire *vs.* attitudes sous contrôle volontaire, que nos recherches se situent.

### II.3.3. Variation inter-individuelle et notion de personnalisation

Nous avons vu dans cette deuxième partie du chapitre que dans une interaction communicative, de nombreux types de comportements expressifs peuvent être observés chez les participants. Ces comportements vont s'inscrire dans les paramètres de la situation et de la tâche, et renseignent sur la personnalité des participants.

Concernant l'adéquation d'un comportement particulier au *sender*, il est nécessaire de garder à l'esprit la grande variabilité inter-individuelle du comportement expressif.

Ainsi, Kaiser, Wehrle, & Schmidt (1998) ont enregistré puis analysé les comportements faciaux de sujets jouant en situation d'IHM. Le principal résultat est une importante variance inter-individuelle dans l'expressivité faciale. Ils ont pu diviser leurs sujets en « un groupe de faible-expression et un groupe de forte expression. À l'intérieur de chaque groupe, les sujets diffèrent selon la variabilité des expressions montrées, *i.e.* certains sujets montrent une variété d'expressions faciales et de répertoires spécifiques à la situation, alors que d'autres montrent un genre de patron d'expressions faciales "typiques" […]. Toutefois, ce manque d'expressivité ne signifie pas que les sujets sont moins (ou pas du tout) impliqués ou affectés émotionnellement. »[42] (*ibid*, p.82)

Selon Cohn (2007, p.10), « les différences individuelles incluent la gamme de réaction pour les affects positifs et négatifs, le temps de réponse, et la probabilité à se conformer aux *display rules*. »[43]. Les facteurs de ces différences sont variés et résident pour lui dans le tempérament, la personnalité, le sexe, l'ordre de naissance, la socialisation et l'arrière-plan culturel de l'individu.

Dans le sens commun, la personnalité correspond à « ce qui constitue la personne, qui la rend psychiquement, intellectuellement et moralement distincte de toutes les autres. »[44]

Matsumoto (2006) appelle « *emotion regulation* » (littéralement « régulation émotionnelle ») la capacité d'un individu à gérer et modifier ses propres réactions émotionnelles pour accomplir un but. Cette régulation est pour lui différente selon les

[42] Citation originale : « a low-expression group and a high-expression group. Within both groups subjects differed according to the variability of the expressions shown, i.e. some subjects showed a variety of facial expressions and situation specific repertories, whereas other subjects showed a kind of "typical" facial expression pattern […]. However, this lack of expressivity does not mean that she is less, or not at all involved or emotionally aroused. »

[43] Citation originale : « Individual differences include reaction range for positive and negative affect, response latency, and probability of conforming to display rules. »

[44] Définition donnée par le TLFI, consulté la dernière fois le 07/01/2011.

cultures, et il tente de montrer qu'au niveau individuel, elle diffère selon les traits de personnalité de l'individu.

Concernant le lien entre personnalité et culture, les psychologues Markus & Kitayama (1998, p. 66) voient la culture et la personnalité comme deux facteurs étroitement liés :

> « La culture et la personnalité sont analysées, de la manière la plus productive, ensemble et en tant que dynamique dans leur constitution mutuelle. »[45]

Ainsi, la personnalité se développe chez l'individu au sein d'une culture et d'un groupe social particulier, et peut influencer le comportement au même titre que ces deux facteurs, mais à un niveau individuel.

Le modèle de personnalité le plus utilisé, en psychologie comme dans les technologies, est le modèle à cinq facteurs (*Five-Factor Model* ou *big five*) de Costa & McCrae (1992). Ce modèle cherche à mesurer et différencier les types de personnalité à partir de leurs caractéristiques comportementales, et ce sur cinq dimensions :

- Ouverture : capacité à changer ses standards dans des nouvelles situations, attitude envers les nouveaux éléments ;
- Conscience : sens du devoir, recherche de réussite, auto-discipline ;
- Extraversion : bonne volonté pour communiquer ;
- Agréabilité : confiance, altruisme, modestie, adaptation aux autres ;
- Névrosisme : anxiété, colère, dépression, impulsivité, vulnérabilité.

Ce modèle est largement implémenté au sein de technologies. Nous relèverons seulement les travaux de Egges, Kshirsagar, & Magnenat-Thalmann (2004) qui ont cherché à établir un lien entre les modèles de personnalité multi-dimensionnels (comme celui à cinq facteurs) et le modèle OCC des émotions (*cf.* Chapitre 1 II.2.).

Par ailleurs, nous avons vu dans le Chapitre 1, parties II.3. et II.4., que la subjectivité est considérée comme une dimension essentielle dans les théories de l'*appraisal*. L'approche socio-cognitive de la psychologie s'est donc également intéressée au lien pouvant exister entre la personnalité et les expressions des émotions. Cette approche part du fait que chaque individu a une personnalité qui lui est propre et qui réside sur des fondements plus ou moins stables concernant ses motivations, ses émotions et ses valeurs. C'est ce qui fait que l'individu répond de façon unique à des événements ou des circonstances particulières : il évalue chaque situation comme plus ou moins importante et pertinente, en fonction des bases de sa personnalité (Wranik, 2009).

Certains traits ou bases de personnalité, faisant partie ou non du modèle à cinq facteurs, sont directement liés à des dimensions d'*appraisal* (*cf*. Chapitre 1, parties II.3.

[45] Citation originale :« Culture and personality are most productively analyzed together as a dynamic of mutual constitution. »

et II.4., et Scherer, Schorr, & Johnstone (2001). Un bon aperçu des différentes études se préoccupant de cette relation peut être trouvé dans Wranik (2009, p. 368-374).

Ainsi, selon Wranik (*ibid*, p. 368) :

> « Il est possible de prédire quelles variables de personnalité influencent systématiquement des dimensions de l'*appraisal* dans des conditions particulières ; »

Notons que le modèle à cinq facteurs, ainsi que les liens établis entre éléments de la personnalité et modèles de l'*appraisal,* s'intéressent à la personnalité intrinsèque du sujet. Quant à nous, nous ne nous préoccupons pas de savoir en quoi la personnalité intrinsèque du sujet influence son comportement. Ce qui nous intéresse concerne la manière dont le sujet est perçu à travers son comportement, quelle personnalité lui est attribué en fonction de ce dernier.

### II.3.4. Notre propre démarche

Nous avons vu dans les deux premières parties de ce chapitre, que dans une interaction communicative, de nombreux comportements expressifs (expressions du *FoT)*, peuvent être observés chez les participants en continu, qu'ils prennent prennent ou non la parole :

- des états mentaux, comme la concentration ou le découragement ;
- des états affectifs, qui se déclinent en attitudes telles que la politesse, en émotions comme la surprise involontaire, ou encore en humeurs, comme le stress ;
- des indices concernant la motivation, les intentions et l'attention du sujet.

Ces indices comportementaux vont s'inscrire dans les paramètres de la situation et de la tâche, et renseignent sur la personnalité des participants. Ainsi, pour être capable de décrire la pertinence d'un comportement expressif, c'est-à-dire son adéquation aux paramètres « quand, où, pourquoi, à qui ,qui » de la situation, nous proposons de décrire l'interaction communicative en tant que processus situé.

Pour cela, il est nécessaire d'étudier un corpus de données permettant d'observer comment différents sujets humains se comportent dans une situation similaire. L'objectif est ainsi de modéliser les variations inter et intra-individuelles observées à travers ces études de cas. Seule une telle démarche permettra par la suite de générer des comportements à la fois pertinents par rapport à la situation et cohérents entre eux.

En effet, les théories des « émotions de base » s'appuient sur des expressions prototypiques. Or, nous avons vu (partie I.3.1.) que les expressions « naturelles », produites spontanément, sont subtiles et de nature complexe, même si nous connaissons certainement les expressions prototypes en tant qu'ensemble de

caractéristiques, et que nous sommes capables de les produire volontairement afin de transmettre un message particulier. Notons que même les expressions prototypiques sont spécifiques à une situation particulière : *e.g.* l'expression prototypique de la colère ne peut être produite avec pertinence dans n'importe quelle situation.

Faisons un parallèle avec d'autres domaines proches, qui ont désormais évolué vers l'intégration de la variation dans les technologies :

- en reconnaissance de la parole était utilisée la comparaison entre des morceaux du signal à analyser, et des ensembles de valeurs formantiques prototypiques (c'est-à-dire mesurés sur des voyelles isolées et bien articulées) pour chaque voyelle. Or les systèmes de reconnaissance étaient alors mauvais, car il est impossible de trouver de telles voyelles dans un flux de parole spontanée. Les systèmes de reconnaissance vocale les plus aboutis modélisent la variation, c'est-à-dire dans quel mesure un son peut potentiellement être perçu comme tel ou tel phonème ;

- la synthèse vocale fonctionnait par des règles, qui concaténaient des segments stéréotypiques, enregistrés dans des phrases porteuses, puis modifiait a posteriori la prosodie. Elle se fait désormais plutôt « par corpus », c'est-à-dire en sélectionnant quels segments correspondent le mieux aux indices linguistiques, prosodiques, stylistiques, dans un large corpus représentatif du « style » prosodique de la situation à modéliser ;

- la reconnaissance des visages utilise aujourd'hui la variation du visage au travers ses mouvements pour fournir une « représentation », une modélisation assez stable de ce visage afin de pouvoir le reconnaître (Bindemann, Burton, Langton, Schweinberger, & Doherty, 2007).

Nous ne chercherons donc pas, comme en psychologie expérimentale ou en linguistique informatique, à cumuler le maximum de sujets et de comportements différents de manière à obtenir des moyennes et des écart-types stables et valides. Nous tenterons d'observer finement, exhaustivement, et à différents niveaux (des différentes modalités, mais aussi à un niveau multimodal et temporel), les comportements d'un nombre restreint de sujets placés dans une situation similaire. Il s'agit d'une méthodologie éprouvée par la phonétique expérimentale.

C'est seulement une fois qu'un certain nombre de phénomènes auront été observés et analysés sur ces sujets (en termes de similitudes et de variations), que nous tâcherons de les valider perceptivement, auprès d'un nombre suffisant de juges naïfs. Ainsi, l'objectif n'est pas de généraliser comment les comportements sont produits par les sujets, mais de généraliser comment les comportements (produits par un humain ou

générés) sont perçus par des « juges » naïfs, en pointant en même temps la variabilité intra- et inter-sujets des comportements.

# III. Cadre théorique : le modèle C-Clone d'Aubergé

Après cette précision sur la démarche adoptée, et plus globalement après cet aperçu des points essentiels concernant le comportement expressif et sa modélisation, nous allons préciser le cadre théorique de l'équipe de recherche au sein de laquelle nos travaux ont été réalisés : le modèle « C-Clone » (pour « Communicative Clone » -Aubergé, Audibert, & Rilliard (2006)-). Ce modèle de communication est général et tient compte des positions théoriques et représentations de notre équipe de recherche. Précisons toutefois que nos travaux sont indépendants de celui-ci. Néanmoins, ils tentent indirectement de développer certaines parties de ce modèle (en particulier en préciser les parties concernant la face et la gestualité du corps, et placer la composante « événements vocaux »). Par conséquent, il permet d'expliquer certaines points de notre problématique, que nous développerons ensuite.

Véronique Aubergé et son équipe de recherche du département Parole et Cognition du GIPSA-lab, ont pour objectif global de modéliser les compétences communicatives de l'humain (à la fois linguistiques et affectives) dans une architecture cognitive d'un agent communicant anthropomorphique appelée « C-Clone ».

Conformément à la distinction établie dans la partie II.3.2. du présent chapitre, cette architecture est à la fois régie dans le temps événementiel (pour la fonction émotionnelle) et dans le temps linguistique (pour les fonctions attitudinale et expressive).

Issu de l'hypothèse cognitive, C-Clone a une architecture modulaire dont l'organisation est coopérative et multi-agents. En effet, ce système, dirigé par les buts de l'interaction verbale, utilise un ensemble de fonctions globalement régies par le système et qui émergent des interactions entre les différents modules (voir Figure 9), dans une architecture complexe, typiquement multi-agents (Aubergé, 2002a).

Chacun d'eux (lexique, morphologie, syntaxe, prosodie, voix, face, gestualité) peut ainsi se définir par une morphologie autonome (mais non indépendante), des degrés de liberté et des contraintes qui lui sont spécifiques. Par ailleurs, pour une fonction communicative donnée, plusieurs modules coopèrent (stratégie intra-fonction), chacun étant un système autonome contrôles/contraintes. Par exemple la fonction de segmentation / hiérarchisation de l'énonciation (qui permet de couper l'énoncé en syntagmes et d'ordonner paradigmatiquement ces derniers) est prise en charge universellement par la syntaxe bien sûr, mais aussi par la prosodie (*e.g.* Aubergé, 2002b), et par la gestualité (McNeill, 1992).

Le choix opéré par l'agent de donner plus d'intelligibilité à une fonction plutôt qu'à une autre (*e.g.* moins d'intelligibilité de segmentation mais plus d'intelligibilité de focalisation) est aussi un vecteur d'information fondamental (stratégie inter-fonctions).

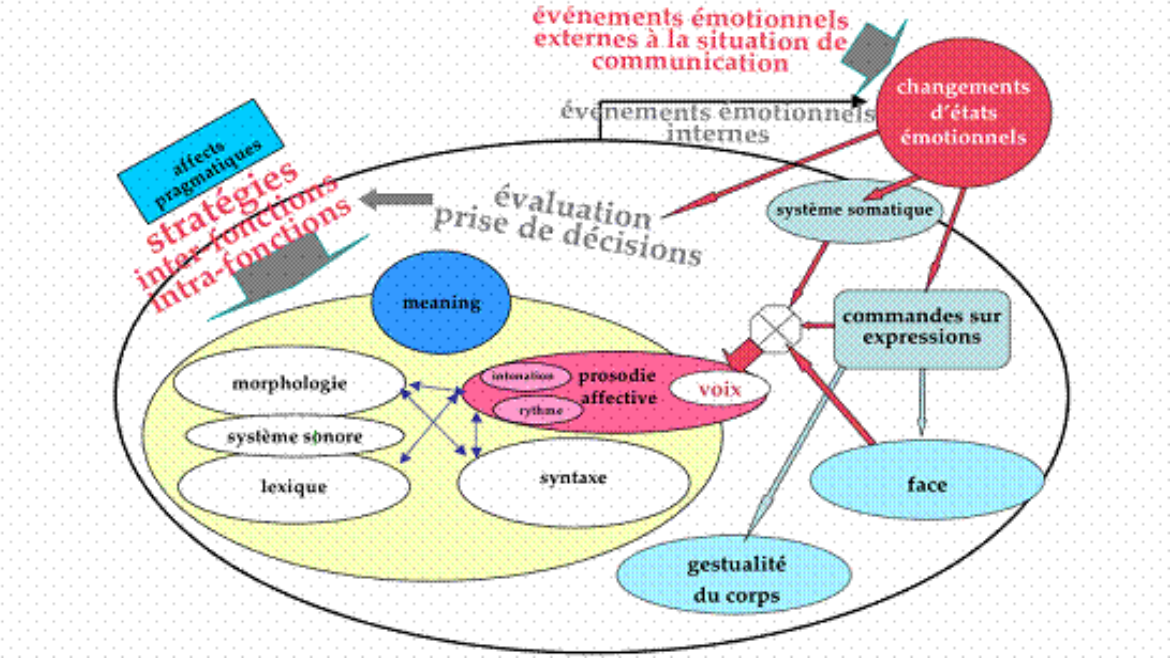



Figure 9: L'architecture cognitive « C-Clone » (Aubergé et al., 2006)

Dans ce modèle, la fonction émotionnelle contrôle par des processus automatiques, les expressions involontaires avec une répartition de l'information dans trois modalités : la voix, la face, le geste. C'est en particulier à travers l'intégration de ces deux dernières dans la représentation interne du modèle, que C-Clone pose le problème de *l'embodiement* (littéralement « corporéisation »).

# IV. Résumé et problématique globale adoptée

L'interaction entre deux humains (ou par analogie restreinte entre un humain et un ordinateur) s'organise autour des tours de parole. Pourtant, le flux de communication de chaque inter-actant est continu : qu'il tienne la parole ou l'écoute, il envoie régulièrement à son interlocuteur des signaux sur ses émotions, intentions ou états mentaux reliés à sa situation de locuteur parlant ou écoutant.

Ainsi, en dehors de son tour de parole et en continu, l'interactant produit des expressions variées. Il exprime en particulier l'évolution de ses processus cognitifs, comme par exemple son état de compréhension ou de résolution mentale d'une tâche, ou ses réactions affectives aux propos du locuteur, humain ou virtuel (comportement de *feedback* et expression du *backchannel*).

Les informations exprimées couvrent un spectre informationnel large et expriment des états mentaux (concentration, *Feeling of Knowing* –*FoK*-), intentions, attitudes (politesse, accord, etc.) et autres états affectifs (stress, déception, irritation, calme, humeurs, etc.). Nous avons regroupé l'expression de ces différents états sous le terme générique de *Feeling of Thinking*, par analogie terminologique au phénomène du *FoK* tel qu'il est étudié par Swerts et al. (2005). Ces états apparaissent typiquement en situation personne-machine, lorsque la personne réagit à une interaction, ou prépare mentalement son interaction (tâche à résoudre, acte de langage, etc.).

Ces expressions semblent ainsi être une trace structurée de la pensée et des états affectifs, et nous souhaitons comprendre de quelle manière. Qu'elles soient volontaires ou involontaires, nous cherchons à les situer dans une multimodalité complexe mettant en complétion, redondance ou succession, les informations qu'elles véhiculent par la voix, le langage, la gestualité et la face. La question sous-jacente est alors de connaître le statut de ces différentes informations, c'est-à-dire s'il s'agit de signaux dédiés ou d'indices utilisés pour l'expression des états relatifs à l'humain communicant. Par exemple certaines personnes « se grattent le nez ». Nous nous demanderons si ce mouvement est caractéristique de la tâche qu'ils sont en train d'effectuer, ou tout simplement lié à leur rhume.

D'autre part, nous nous intéresserons à l'organisation temporelle de ces indices / signaux : lorsque cette même personne se gratte le nez plus fort et plus souvent lorsqu'elle se trouve dans un état spécifique (*e.g.* un état de gêne), ce mouvement pourrait être un indice indirect de cet état.

Étudier l'expression de ces informations avec une telle démarche pourrait à la fois apporter de nouvelles connaissances sur le comportement humain, et donner de nouvelles idées aux concepteurs d'agent virtuels communicants.

En effet, au niveau technologique, modéliser et corporéiser les expressions émotionnelles, et plus globalement le comportement expressif, apparaissent bénéfiques aux applications dans la majorité des cas. C'est pourquoi le domaine est en pleine expansion. Dans le but de rendre les ACAs crédibles, il est nécessaire de modéliser des comportements expressifs de locuteurs mais aussi d'auditeurs. Ces comportements doivent être considérés dans leur multimodalité, être pertinents par rapport à la situation de communication (au sens large), et former un ensemble cohérent pour chaque agent. Le but est non pas de trouver et de générer un comportement moyen, standard, mais de modéliser la variation comportementale au sein d'une communication située. Nos travaux consisteront donc en des études approfondies de cas, sur des sujets à la personnalité perçue, et par conséquent aux comportements expressifs et singuliers. Ainsi, nous chercherons à associer différentes variations avec leurs facteurs situationnels et inter-individuels.

Pour mener à bien ces travaux, nous observerons et analyserons des données naturelles prises dans des situations où l'agent aura à interagir avec l'humain. Nous allons développer dans le chapitre suivant les différents enjeux méthodologiques liés à une telle démarche.

# CHAPITRE 3 : ENJEUX MÉTHODOLOGIQUES LORS DE L'ÉTUDE DE L'EXPRESSION DU *FEELING OF THINKING*

Outre les enjeux théoriques et technologiques inhérents à l'étude des expressions de ce que nous avons appelé *Feeling of Thinking*, l'étude de ces phénomènes passe par l'observation de données réelles, c'est-à-dire de corpus. Un certain nombre de problématiques apparaissent lorsque l'attention est portée sur la méthodologie à utiliser pour décrire les formes des expressions et annoter leur valeur, mais aussi dans un premier temps pour élaborer ou choisir ce corpus.

En parallèle, deux approches peuvent être adoptées pour étudier le comportement expressif : les tâches d'annotation et de description peuvent être guidées par une théorie, ou au contraire guidées par les données dans une approche inductive.

Nous allons décrire dans ce chapitre les principaux enjeux méthodologiques et les risques qu'ils impliquent, en nous appuyant sur certains courants philosophiques et sur différents exemples de modèles qui nous paraissent caractéristiques. Nous aborderons ensuite la manière dont le domaine de l'éthologie, entre autres, a trouvé réponse à ces enjeux par une méthodologie particulière. Finalement, la méthodologie que nous avons adoptée, inspirée de celle de l'éthologie, sera décrite globalement, puis de manière plus précise et plus technique concernant l'étiquetage, qui est une étape clé pour l'objectivation de notre étude.

# I. Étudier l'humain et son comportement : apports provenant de la philosophie

## *I.1. De l'épistémologie à la phénoménologie de Merleau-Ponty*

Dans la tradition philosophique francophone, l'épistémologie est une branche de la philosophie des sciences. Elle étudie de manière critique les méthodes scientifiques utilisées pour arriver aux théories et résultats, afin de déterminer l'origine, la valeur et la portée objective de ces derniers, c'est-à-dire, en première approximation et selon Piaget, « la constitution des connaissances valables » (cité par Le Moigne, 1995).

Dans le *Discours de la méthode* (1637), Descartes montre l'importance de partir de l'étude des objets les plus simples pour atteindre la connaissance des objets les plus complexes. Si cette méthode fut largement admise et appliquée, la question de la validité de l'empirisme souleva de nombreux débats, en particulier avec l'apparition de l'épistémologie Kantienne.

Kant (1781) apporta en effet un changement radical de point de vue sur l'empirisme, qui jusqu'alors était négligé : pour lui, « aucune de nos connaissances ne précède donc en nous l'expérience ; toutes commencent avec elle » (p.32, introduction), et « il n'y a pas d'autres objets que ceux des sens qui puissent nous être donnés, et ils ne peuvent l'être que dans le contexte d'une expérience possible » (p.182, chapitre 3, section 2).

Cette conception reste actuelle, et d'autant plus dans l'étude des phénomènes affectifs puisque nous n'avons accès aux émotions et autres états d'un locuteur que par la perception de leurs manifestations, via les différentes modalités, et sous forme d'indice comme de signal (*cf.* chapitre 2).

À la suite de Kant, la phénoménologie de Husserl a continué à faire prévaloir les démarches empiriques et la perception, en portant une attention particulière sur l'objectivité. La phénoménologie est, littéralement, la science des « phénomènes » au sens philosophique du terme, c'est-à-dire des choses telles qu'elles se présentent à la conscience et qui sont susceptibles d'acquérir une valeur objective par la répétition ou la reproduction :

> « ce qui apparaît, ce qui se manifeste aux sens ou à la conscience, […], et qui peut devenir l'objet d'un savoir », mais aussi « ce que l'on observe ou constate par l'expérience et qui est susceptible de se répéter ou d'être reproduit et d'acquérir une valeur objective, universelle » (Définitions du TLFI, consulté le 07/09/2010)

Nous retiendrons de cette phénoménologie l'idée qu'un savoir objectif peut naître de ce qui est perçu dans certaines conditions.

La philosophie de Husserl a, entre autres, été développée par la suite par le philosophe français Merleau-Ponty, dont les travaux sont partis de l'étude de la perception.

Pour lui, la perception a une dimension active en tant qu'ouverture primordiale au monde vécu, au « LebensWelt » (Merleau-Ponty, 1990, p.235-236). De plus, il a développé la notion de corporéité, considérant que le corps est une condition permanente de l'expérience et qu'il constitue l'ouverture perceptive au monde, ce dont l'analyse de la perception doit tenir compte. Notons que cette idée va à l'encontre de la dualité corps-esprit de Descartes, et a eu un impact sur les théories actuelles de la psychologie.

Outre l'importance de considérer la manière de percevoir les objets par le biais du corps, nous retiendrons également de cette corporéité la dimension d'expressivité qui en découle, et les questions qu'elles posent :

> « Qu'il s'agisse des vestiges ou du corps d'autrui, la question est de savoir comment un objet dans l'espace peut devenir la trace parlante d'une existence, comment inversement une intention, une pensée, un projet peuvent se détacher du sujet personnel et devenir visibles hors de lui dans son corps, dans le milieu qu'il se construit. » (Merleau-Ponty, 1976, p. 401).

## *I.2. Lévi-Strauss ou les idées non conformistes d'un ethnologue*

> « C'est dans l'homme même qu'il faut étudier l'homme : il ne s'agit pas d'imaginer ce que l'homme aurait pu ou dû faire, mais de regarder ce qu'il fait » (maxime du président de Brosses, souvent citée par Lévi-Strauss)

Lévi-Strauss est un ethnologue français du XXème siècle. Une de ses grandes contributions est d'avoir introduit la méthode structurale[46], déjà utilisée par les Saussuriens en linguistique, dans le domaine de l'anthropologie (Lévi-Strauss, 1958). Plus globalement, toute son œuvre porte sur les relations entre le sensible et l'intelligible. Il a notamment critiqué la méthode souvent utilisée par la sémiologie, malgré son accord avec l'idée de l'importance du signifiant qui découle, pour lui, de la supposition que les énergies psychiques sont codées :

> « L'erreur d'une certaine sémiologie, ce fut de prendre les choses à l'envers : de ne voir dans les données que des signes et de tenter souvent artificiellement de leur conférer une cohérence, alors qu'il s'agissait d'identifier dans le matériau observé les éléments dont les valeurs différentielles en forment la trame logique. » (Levi-strauss, cité par Hénaff, 2009, p.22-23)

[46] La méthode structurale « isole un ensemble d'éléments et de relations formelles pour étudier un phénomène, sans faire appel à la signification » (définition du TLFI, consulté le 09/09/2010)

Lévi-Strauss s'est également intéressé aux inter-relations qui peuvent exister entre la langue et la culture, qui représentent pour lui « deux modalités d'une activité plus fondamentale, l'activité de "l'esprit humain" ». Sa particularité est ici sa perspective naturaliste : il préconise de « réintégrer la culture dans la nature » (Lévi-Strauss, 1962, cité par Hénaff, 2009, p.23).

Selon Sperber (2009), il est un des précurseurs des idées de la « révolution cognitive », qui aborde l'esprit humain de façon naturaliste et a ramené la psychologie à l'étude des mécanismes de la pensée. De plus, comme il l'avait prédit, les psychologues cogniticiens se sont rendus compte ces dernières années que « les structures mentales se révèlent non seulement dans les expériences de laboratoire, mais aussi dans leur manifestation culturelle ». Ainsi, le « relativisme culturel » est pour lui un principe méthodologique qui oblige toute étude d'ethnologie à adopter le mode du « regard éloigné », c'est-à-dire « d'oublier qui l'on est pour comprendre qui sont les autres » (Lévi-Strauss, 1988, cité par Sperber, 2009, p.24).

Lévi-Strauss a donc fortement influencé la recherche en sciences humaines et sociales par ses considérations sur la culture, sur son implication dans les manifestations des « structures mentales », et de son influence sur la perception. Ces idées apparaissent également dans les domaines de l'ethnométhodologie et de l'éthologie humaine (voir la partie III. de ce chapitre).

### *I.3. Empirisme, induction et méthodologie expérimentale*

L'empirisme postule que toute connaissance provient essentiellement de l'expérience. Représenté par exemple par les philosophes Bacon (1561-1626), Locke (1632-1704) et Hume (1711-1776), ce courant considère que la connaissance se fonde sur l'accumulation d'observations et de faits mesurables, dont sont extraites des lois générales par un raisonnement inductif. L'induction consiste à se fonder sur l'observation de cas singuliers pour justifier une théorie générale. L'induction s'oppose à la déduction logique, qui se fonde sur des axiomes ou des définitions, et est à la base de la méthode expérimentale.

Alors que l'empirisme est le fait d'apprendre par l'observation et par l'expérience, l'expérimentation part de la théorie. Elle teste ensuite par des expériences répétées la validité d'une hypothèse et cherche à obtenir des données quantitatives, de manière à pouvoir affiner la théorie, à l'ajuster aux résultats, aux observations. Les principes de l'expérimentation sont que les lois sont hypothético-déductibles, et que les expériences doivent être à la fois reproductibles, vérifiables et falsifiables (*cf.* Popper, 1972).

Le travers méthodologique existant dans l'expérimentation est que cette dernière ne produit que des résultats tautologiques, c'est-à-dire déjà inscrits dans les prémisses. Il risque d'en découler un biais : que les données testées pour évaluer la théorie soient choisies pour que la théorie fonctionne. La valeur des résultats est donc fonction de la rigueur avec laquelle ils ont été obtenus.

Une conséquence est que la déduction logique, et donc l'expérimentation, ne produit aucune nouvelle connaissance, puisqu'elle est analytique et que ce qui est testé est contenu dans la théorie. À l'inverse, l'induction est susceptible d'enrichir les connaissances par de nouvelles propositions qui émanent des données : elle est synthétique.

Malgré cette mise en avant de l'empirisme par les idées philosophiques décrites dans les parties précédentes, la méthode expérimentale lui est complémentaire et garde un large intérêt, surtout dans un second temps : elle permet de valider scientifiquement, quantitativement, des observations dégagées empiriquement.

Pour éviter les travers de l'expérimentation, un compromis peut être :

- d'adopter dans un premier temps une méthodologie empirique sur des données naturelles, en évitant les *a priori* sur ce qui va être trouvé ;
- puis de soumettre à l'expérimentation certains phénomènes observés de manière à construire un modèle, ou du moins préciser voire remettre en question certaines parties de modèles déjà existants.

Il sera alors estimé que la proposition théorique issue des observations est valable si l'analyse et / ou les résultats de l'expérimentation sont reproductibles.

# II. Enjeux méthodologiques des modèles actuels et présentation du corpus

Avant de déterminer la méthodologie de notre propre étude, qui se veut inductive, faire un survol des modèles actuels de gestualité et d'expression des affects nous aidera à établir les enjeux méthodologiques de ce type d'étude.

En effet, un modèle particulier implique toujours une méthodologie particulière : le modèle utilisé influence l'analyse en agissant comme un filtre perceptif. Le chercheur se focalise alors sur certains éléments (son objet d'étude), et peut en oublier certains paramètres, ou même certains objets, pourtant susceptibles d'être pertinents pour l'interaction communicative et ses expressions affectives.

C'est pourquoi il est utile de donner un aperçu de ces modèles en adoptant un regard critique sur leur méthodologie et ce qui en découle. Nous nous attacherons ainsi à déterminer les contraintes qu'ils apportent éventuellement à la perception et à l'analyse des données. Nous relèverons également les restrictions qu'ils imposent à leur objet d'étude, et qui peuvent les empêcher d'avoir une vision globale du comportement expressif.

D'un point de vue méthodologique, les paradigmes d'études de perception des expressions faciales et des émotions se sont transformés au cours des dernières décennies. À l'origine et pendant de nombreuses années, étudier la perception visuelle des émotions consistait généralement à demander à des juges d'associer des photographies stéréotypées d'expressions d'états émotionnels prototypiques (le plus souvent six émotions, dites de base par Ekman en particulier) aux labels correspondant à ces dernières. Les recherches en *Affective Computing* ont changé ce paradigme afin de faire face à l'enjeu technologique du réalisme (Cowie, 2009). En effet, « décrire la manière qu'ont les émotions de "teinter" actions en cours et interactions est un problème différent de celui de catégoriser de brefs épisodes d'émotions relativement "pures". Ainsi, décrire cette "teinture émotionnelle" est un défi en lui-même »[47] (*ibid*, p. 3515).

Avec l'essor des travaux de l'*Affective Computing* sont donc apparus de nouveaux enjeux théoriques et méthodologiques : chercher à modéliser l'expression des émotions et autres états affectifs a révélé la manière dont cette capacité à percevoir et déterminer les émotions des autres, ordinaire pour les humains, est d'une complexité extraordinaire (Cowie, 2009).

[47] Citation originale : « describing that colouring is a different problem from categorizing brief episodes of relatively pure emotion. […] Describing emotional colouring is a challenge in itself ».

## *II.1.Enjeux du recueil de corpus*

Nous avons vu dans le Chapitre 2, partie II.3., que l'expression émotionnelle diffère selon qu'elle est simulée, actée, prototypique, ou au contraire produite spontanément, mais aussi que le comportement expressif « naturel » était de nature complexe. Ces considérations impliquent que « si notre but est de construire un système d'Interaction Homme-Machine "non-caricatural" émotionnellement, nous devons nous focaliser sur des bases de données authentiques et sur l'interaction naturelle avec des émotions issues de causes authentiques, plutôt que sur des données biaisées avec des événements artificiels »[48] (Campbell et al., 2006, p.xxv).

De plus, des études menées en particulier par Ekman (2003) et Matsumoto (Matsumoto, Yoo Hee, & Fontaine, 2008) ont montré que le comportement était modifié en présence d'autres personnes, selon des règles dépendantes entre autres de la culture : c'est ce qu'ils appellent les *display rules* et qui peuvent se traduire par une inhibition ou encore une exagération du comportement (Chapitre 2 II.3.).

Il est donc nécessaire de travailler sur des corpus présentant des expressions naturelles, c'est-à-dire à la fois spontanées et susceptibles d'apparaître en situation réelle. De plus, l'adjectif « spontané » peut ici être compris en opposition à « acté, simulé », mais aussi comme « en évitant l'inhibition, sans influence de l'observateur ».

En effet, par le fait même de vouloir observer, l'observateur modifie le comportement du sujet, et cela biaise le corpus, d'autant plus que l'intérêt est porté sur l'état affectif du sujet et son expression. Nous sommes donc confrontés à un paradoxe du même type que le « paradoxe de l'observateur », introduit par le sociolinguiste Labov (1973), et développé par Pooley (1996, p.80) :

> « Une attention considérable doit être porté aux dangers du "paradoxe de l'observateur", c'est-à-dire le fait que la présence de l'observateur peut détruire le phénomène qu'il observe. »[49]

Les méthodes d'établissement de corpus émotionnels peuvent être classées selon trois axes :

- *in vivo vs. in vitro* : les méthodes *in vitro* renvoient aux enregistrements en laboratoire, qui permettent des conditions d'enregistrement optimales ; les autres sont des enregistrements sur le terrain ;

[48] Citation originale : « If our goal is to build emotionally 'non-caricatural' Human-Machine Interaction system, we have to focus on real-life databases and natural interaction with genuine emotional cause events instead of on biased data with artificial events ».

[49] Citation originale : « Considerable care was taken to avoid the dangers of the 'observer's paradox', *i.e.* the fact that the presence of the observer may destroy the phenomenon that s/he is observing ».

- le degré de contrôle du contenu et de la situation de l'enregistrement ;
- les protocoles actés *vs.* authentiques, avec les niveaux intermédiaires de l'induction et de l'élicitation.

Bien que ces trois axes soient totalement indépendants, certains paramètres sont toutefois liés. Par exemple des données authentiques sont plus facilement recueillies *in vivo* et sans contrôle de l'observateur. Des états de l'art sur le recueil de corpus émotionnels ont été établis par Campbell (2000), Douglas-Cowie, Campbell, Cowie, & Roach (2003) ou encore Scherer (2003) en ce qui concerne les expressions vocales des émotions.

L'objectif est dans l'idéal de travailler sur un corpus *in vivo* d'expressions authentiques et spontanées, mais avec des enregistrements de la même qualité qu'en laboratoire et un degré de contrôle sur le contenu élevé. Différentes stratégies ont donc été élaborées.

Pendant longtemps, des acteurs ont été utilisés dans le cas des corpus *in vitro*, avec des techniques plus ou moins élaborées d'élicitation, qui consistent à demander aux sujets de se replonger dans une situation chargée émotionnellement et qu'ils ont réellement vécue (*e.g.* Mozziconacci, 1998). En effet, selon Bänziger & Scherer (2007), le fait d'encourager les acteurs à réactiver des expériences émotionnelles passées leur permettrait de produire plus facilement des expressions similaires à des expressions spontanées.

De l'autre côté, différentes techniques d'induction chez des locuteurs non professionnels ont également été utilisées : soit des stimuli à fort contenu émotionnel sont présentés juste avant la production de parole (*e.g.* Tolkmitt & Scherer, 1986) ; soit la tâche consiste en la lecture de textes fortement chargés sur le plan émotionnel (*e.g.* Wilting, Krahmer, & Swerts, 2006, en adaptant la méthode d'induction de Velten, 1968).

Cependant, ces techniques *in vitro* ont l'inconvénient de rendre les sujets centrés uniquement sur eux-mêmes, sur leur vécu ou sur leur imagination ; l'interactivité et la spontanéité ne sont donc pas présentes.

Prendre en considération ce point est d'autant plus important que des études montrent que la parole actée est spécifique par divers aspects. Une première expérience menée par Aubergé & Cathiard (2003) a par exemple montré que l'amusement acté pouvait être discriminé de l'amusement non acté.

En effet, les émotions exprimées par des acteurs ont une naturalité dépendante à la fois du contexte artistique (théâtre, cinéma, improvisation), des méthodes et des

cultures. Ainsi, elles peuvent être fort éloignées d'une imitation ou simulation d'expressions authentiques.

Le fait que ces productions soient aisément identifiables dans des tests perceptifs, ne signifie pas qu'elles soient identiques à une production authentique. Une parole actée très caricaturale et stéréotypée peut assurément donner de meilleurs scores d'identification que de la parole authentique. D'autre part, nous ne rangeons pas les énoncés simulés sous le terme émotions, puisqu'il s'agit alors de processus intentionnels. Enfin, lorsque la parole est produite par un acteur dont la méthode est de « re-ressentir » une émotion vécue, la compétence intrinsèque de l'acteur et la qualité de ses performances en laboratoire (situation qui ne favorise pas son jeu) ne permettent ni de maîtriser, ni d'évaluer la qualité d'imitation des productions.

Par conséquent, d'autres études, moins répandues, ont cherché à recueillir des expressions authentiques tout en conservant le degré de contrôle élevé propre aux enregistrements *in vitro.* Ce recueil est rendu possible par la présence d'un compère, dit magicien d'Oz, associée à l'utilisation de tâches précises permettant de contraindre la nature des énoncés recueillis et d'induire les variations émotionnelles attendues.

De cette manière Johnstone, van Reekum, Hird, Kirsner, & Scherer (2005) ont utilisé un scénario de jeu vidéo, et (Kaiser, Wehrle, & Schmidt, 1998) ont imaginé un scénario d'interaction fondé sur une tâche informatique. Quant à nous, afin d'obtenir un corpus d'expressions émotionnelles authentiques, nous avons mis nos sujets en pseudo-interaction avec la machine, dans une application ayant un scénario précis, en ne leur donnant qu'un langage de commande vocale réduit et en ne leur permettant pas d'influer sur le déroulement du scénario (*cf.* partie III.2.1. de ce chapitre).

## *II.2.Notre corpus : E-Wiz SoundTeacher*

### II.2.1. Contraintes

Notre problématique nous impose de travailler sur un corpus multimodal d'expressions spontanées. De plus, il s'agit de contrôler son contenu verbal afin de faciliter l'exploitation des données, et de mettre des sujets dans une même situation afin d'avoir des données comparables entre elles.

Afin d'avoir la possibilité d'un contrôle suffisant, notre corpus a été recueilli en situation d'interaction personne-machine. Dans le but de récolter des expressions liées à des réactions émotionnelles assez proches (dépendantes uniquement du profil psychologique du sujet), la tâche et les sujets ont été choisis de manière à ce que ces derniers s'impliquent et soient motivés par la tâche elle-même, et non par des motivations indirectes (rétribution, note, etc). D'autre part, afin d'obtenir un corpus

d'expressions émotionnelles authentiques tout en gelant l'expressivité verbale (en dehors de commentaires libres), les sujets ont été mis en pseudo-interaction avec la machine, dans une application ayant un scénario précis. Il ne leur a été donné qu'un langage de commandes vocales réduit, et les sujets ne pouvaient influer sur le déroulement du scénario.

### II.2.2. Description du corpus

Le corpus que nous utilisons pour nos travaux a été recueilli par Nicolas Audibert, Véronique Aubergé et Albert Rilliard, en utilisant le scénario « Sound Teacher » sur la plate-forme E-Wiz. Pour plus de détails sur ce corpus et son recueil, se référer à Aubergé, Audibert & Rilliard (2006), ou Audibert (2008).

Ces données d'interaction personne/machine consistent en des enregistrements de dix-sept sujets, piégées dans un même scénario d'induction émotionnelle. Pourtant, de par leur profil psychologique, ils présentent des réactions différentes en termes d'états mentaux et affectifs.

Ces sujets, 11 femmes et 6 hommes, sont vus de face à partir de leur buste. Ils sont isolés en chambre sourde face à un écran, et ne se savent pas enregistrés. Ils sont placés dans une situation d'apprentissage des voyelles des langues du monde, utilisant un pseudo système révolutionnaire : « Sound Teacher ».

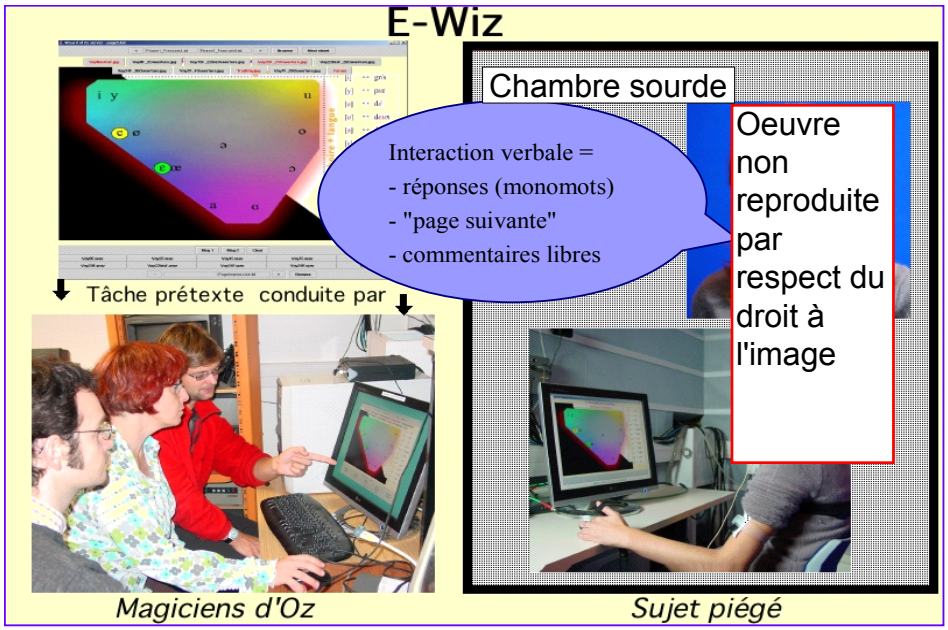



Figure 10: Le corpus E-Wiz SoundTeacher et le paradigme du Magicien d'Oz

D'un côté, les sujets interagissent seulement par la parole, pour fournir les réponses ou donner des commentaires libres (pas de clavier ni souris) ; de l'autre, la machine « répond » soit par du texte, soit par l'exécution de la tâche demandée.

Ainsi, ce corpus a l'avantage de fournir des expressions émotionnelles authentiques mais contrôlées grâce à l'utilisation du paradigme du magicien d'Oz : le sujet pensait communiquer avec un ordinateur, alors qu'en réalité, le comportement apparent de l'application était géré à distance par le magicien (voir Figure 10).

Le scénario peut être découpé en quatre grandes phases, déterminées par le type de *feedback* fourni au sujet sur ses résultats (ce dernier étant manipulé par le magicien pendant toute la durée du scénario).

La première phase, dite d'entraînement, familiarise et rassure le sujet. Une deuxième phase implique le sujet dans des tâches très simples pour lesquelles il est félicité. Cela a globalement induit chez l'ensemble des sujets des émotions positives. Les troisième et quatrième phases, de complexité croissante, renvoient au sujet des *feedbacks* négatifs, la quatrième et dernière phase étant la répétition biaisée de la tâche initiale simple, retournant au sujet des résultats cette fois très mauvais, dans le but de l'inquiéter, de le stresser, ou du moins de le déstabiliser.

Parallèlement et simultanément à l'enregistrement vidéo, il a été procédé à l'enregistrement de différents paramètres physiologiques pour chaque sujet, mesurés par les capteurs biophysiologiques non intrusifs d'un équipement Procomp : conductance et température de la peau, fréquence cardiaque, rythme respiratoire, EMG (électromyogramme). L'étude des variations physiologiques des sujets pourrait permettre à plus long terme, de mieux suivre l'évolution cognitive / affective du sujet en parallèle de l'organisation temporelle scénario de l'interaction personne-machine.

Après chaque enregistrement, le sujet, après s'être fait expliqué la manipulation dont il a été victime, a auto-annoté sa production en notant par la modalité de son choix (langage écrit, dessins, signes, etc.) ses états mentaux et affectifs, au fil de l'avancement du scénario. Il ne leur a pas été demandé explicitement leurs états émotionnels, car les sujets auraient pu avoir tendance à répondre comment ils auraient aimé être (et non comment ils étaient). Cette dernière remarque restant d'ailleurs valable, il nous faudra la garder à l'esprit. Nous reviendrons sur les auto-annotations et leur intérêt lors dans la partie III.3.1. de ce chapitre.

Les labels d'auto-annotation naïfs sont destinées à être une monnaie d'échange « naturelle » lors de tests perceptifs auprès de juges naïfs. Ils ont été utilisés en particulier lors de la validation perceptive de l'étiquetage du signal audio-visuel.

### II.2.3. Intérêt du corpus pour l'étude de l'expression du *FoT*

Étant donné le scénario, les sujets sont, lors de la tâche, alternativement en train de lire, de réfléchir ou de produire de la parole. Leurs productions de parole sont

spontanées, tout en restant contrôlées grâce au prétexte des commandes vocales destinées à une machine. Elles peuvent être :

- des réponses monosyllabiques isolées (les couleurs « rouge », « vert », « jaune », « brique » et « sable ») ;
- l'expression « page suivante » ;
- des commentaires libres.

Les sujets utilisent tous le même matériel sonore. Il s'agissait d'un enjeu important pour les études de notre équipe de recherche portant sur la parole (en particulier Audibert et ses travaux sur la prosodie expressive (2008).

« Sound Teacher » est une situation de dialogue minimal : le sujet sait que ce qu'il affiche de ses affects et de ses états mentaux ne changera ni la nature de l'interaction, ni ses relations à l'interactant, puisqu'il se trouve face à une machine (d'autant moins qu'il ne sait pas qu'il est filmé dans le but de futures études). L'interaction n'est pas humanisée (la tâche est directe) et le sujet n'a donc aucune raison d'envoyer des expressions destinées à influencer les états de l'ordinateur. De plus, la phase de communication humaine ou humanisée dans laquelle le sujet auditeur envoie un feedback à son interlocuteur « intentionnellement » n'est donc pas attendue.

Ainsi, lors des enregistrements, le sujet humain était supposé produire des expressions d'états mentaux, d'évolution de processus cognitif, d'humeurs et d'émotions, à l'inverse des expressions sociales, d'attitudes ou d'intentions, *a priori* gelées par la situation. Pourtant ces dernières expressions sont présentes et décrites dans les auto-annotions des sujets eux-mêmes.

Cette situation classique de dialogue homme-machine interactive dans laquelle, pour le sujet, tout ce qui n'est pas verbal n'est pas pris en compte, nous a donc permis de recueillir une gestualité du corps et de la face, ainsi que d'autres événements vocaux de nature diverse, tous spontanés et sans aucune intentionnalité de la part du sujet. En parallèle, un autre intérêt de la situation fournie par « Sound Teacher », est qu'elle n'implique pas (ou peu) d'inhibitions d'expressions du *FoT* par les sujets (en supposant qu'elles puissent être inhibées).

Par ailleurs, nous avons eu la chance d'enregistrer des sujets dont la motivation était similaire, et dans une même situation inductrice d'affectivité. Toutefois, il ne faut pas négliger la seule différence entre les sujets : leur profil psychologique, et plus particulièrement leur personnalité. C'est d'elle que découlent leurs réactions émotionnelles et leur réaction globale face au scénario. Ainsi, il ne sera jamais possible

d'obtenir une uniformité dans les comportements des sujets, et c'est justement l'objet de cette thèse que d'étudier la variabilité des comportements face à une tâche et un scénario identique.

En l'occurrence, alors que certains sujets ont remis en cause leurs propres capacités lors de leur baisse de performance, d'autres, à l'inverse, ont remis en question le système. Il s'agit donc d'une variable à prendre en compte, et il nous sera nécessaire d'établir une typologie basique des différents profils psychologiques rencontrés chez les sujets. Nous reviendrons sur cet aspect au cours des chapitre 5 et 6.

### II.2.4. Adaptation de la notion de tour de parole à notre corpus

Afin d'adapter à la situation d'IHM de notre corpus la notion pragmatique de tour de parole (*cf.* Chapitre 2 I.4.1.), nous la distinguons des « prises de parole ». Plus précisément, une interjection peut être une prise de parole hors tour de parole, de la même manière qu'un *feedback* produit par le sujet lors de la lecture de quelque-chose sur l'écran avec une production lexicalisée (lorsque le sujet « parle tout seul », se parle à lui-même). Dans ce cadre, un tour de parole du sujet humain est une prise de parole « attendue » par le logiciel, c'est-à-dire une réponse ou une commande (*e.g.* « page suivante » à la suite desquelles le logiciel va envoyer soit une réponse en terme de résultats, de nouvelles instructions, ou une nouveau stimulus, soit va exécuter la commande. En quelque sorte, en IHM et en particulier dans notre scénario, une commande vocale ou une réponse du sujet est un signal de *backchannel* « disant » à la machine « c'est ton tour de parler » (ou plutôt d'agir). Cette précision terminologique restera valide et sera utilisée lors de nos analyses (*cf.* en particulier les Chapitre 6 et 7).

## *II.3. Enjeux du codage*

### II.3.1. Présentation des grandes approches et de la problématique

Lorsqu'il s'agit de mesurer les expressions faciales, deux grandes approches méthodologiques sont utilisées (Kaiser, Wehrle, & Schenkel, 2009 p.83) : la première, nommée « méthode des jugements », consiste à utiliser le *degré d'accord inter-juges* pour valider la mesure des expressions ; la seconde, nommée « codage des signes » cherche à « mesurer le comportement facial lui-même par l'observation très exacte de chaque indice facial », les expressions étant notées et classifiées selon des critères prédéterminés.

Bien que ces deux méthodes paraissent proches puisqu'elles utilisent toutes deux des observateurs, une grande différence existe au niveau de ce travail d'observation. Dans la méthode des jugements, les observateurs « font des *inférences* au sujet *des dessous* du

comportement (émotion, humeur, traits, attitudes, personnalité, etc.) » (*ibid*), alors que dans le codage des signes, les observateurs « *décrivent* le comportement lui-même (le nombre de mouvements, leur durée, les muscles mis en jeu, etc.) ainsi que ce qui diffère sur les visages des différentes personnes des groupes » (*ibid*). Cette différence renvoie à l'opposition existant entre « juger » et « coder ».

Davantage d'études sont réalisées en utilisant l'accord inter-juges plutôt que la méthode de codage des signes, pour une question de rapidité et de facilité d'obtention des résultats.

Cependant, comme nous l'avons relevé dans le Chapitre 2, le comportement facial expressif est multi-fonctionnel, et les indicateurs de processus émotionnels peuvent être très subtils et peuvent changer très rapidement. C'est pourquoi il est nécessaire d'utiliser des approches qui permettent de mesurer les expressions faciales de manière objective, sans jugement sur leur signification potentielle, et à un niveau micro-analytique. Il s'agit donc ici de « coder », et non de « juger ».

La question est alors de savoir quoi coder.

De nombreuses grilles (ou schémas) d'annotations, appelées également « schéma de codage », ou encore « représentation des transcriptions » en analyse conversationnelle, ont été développées. Nous citerons par exemple :

- pour les gestes : Kendon (1993), McNeill (1992), Calbris (1990) ;
- pour les expressions faciales : Ekman & Friesen (1978) ;
- pour le regard : Poggi, Pelachaud, & De Rosis (2000) ;
- spécifiquement pour les émotions : Cowie et al. (2000), Ekman (1999b), Scherer (2000) ;
- d'une manière plus globale pour le comportement multimodal : Martin et al. (2006), Blache et al. (2010), Colletta, Kunene, Venouil, Kaufmann, & Simon (2009).

Concernant spécifiquement les différentes représentations utilisées en analyse conversationnelle, un état de l'art se trouve par exemple dans Ochs (1979).

Ces grilles d'annotations diffèrent les unes des autres par les éléments et les paramètres codés, mais également par le(s) niveau(x) d'annotation, ou la granularité du codage. Ces différences relèvent de l'approche voire de l'objectif adopté pour l'étude.

La conception de cette grille d'annotation est d'une importance fondamentale pour la suite des recherches. Le contenu de cette grille ainsi que la manière de noter les comportements sont susceptibles d'influencer considérablement les résultats qui seront obtenus.

Un risque pouvant avoir de lourdes conséquences, est de décider à l'avance, involontairement, par la conception même de la grille, de ce que nous projetons de / allons trouver. Ce risque est d'autant plus grand que l'étude des données cherche à valider un modèle déjà construit.

Ainsi, la grille adoptée filtre l'information, formate la manière de voir les choses, et peut par conséquent faire passer à côté d'informations potentiellement pertinentes. Métaphoriquement, la grille focalise une lumière sur l'objet recherché, les informations pertinentes restant alors dans l'ombre. Il s'agit d'une conséquence liée à la restriction (sous n'importe quelle forme) de l'objet d'étude.

> « Une erreur importante à éviter est l'omission de certains signaux dans le codage, car ils sont prématurément considérés comme non significatifs. »[50] (Hager, 1983)

Dans la suite de cette partie, nous allons détailler trois types d'exemples de recherches, qui nous semblent chacune représentative d'une manière particulière de restreindre l'objet d'étude.

### II.3.2. La description multimodale vue par les interactionnistes et l'analyse conversationnelle

L'analyse conversationnelle s'inscrit dans une tradition interprétative (*cf.* par exemple Mondada, 2008). La description concerne uniquement ce qui fait sens à travers la situation, à travers la réaction effective des autres participants. Le critère des interactionnistes est que tout ce qui est transcrit au niveau non-verbal doit pouvoir être justifié au niveau analytique, en termes d'interaction communicative sociale. Ainsi, un choix *a priori* est effectué sur ce qui est pertinent.

De la même manière, la granularité d'annotation est déterminée par cette pertinence *a priori* des objets. Par exemple lors de la description d'une situation de jeu de cartes, l'emploi du terme « piocher » ne pose pas de problème tant que l'action de piocher n'est pas un objet d'étude de l'analyse. Si cela est déjà le cas ou si ça le devient, la granularité pour décrire ce type d'action devra être plus fine/détaillée, et plus objective. Il s'agit de la question du focus d'analyse, découlant du fait qu'en analyse conversationnelle, il n'est pas possible de tout décrire exhaustivement. Toutefois, les interactionnistes évitent d'interpréter les comportements en termes d'intentionnalité.

Par ailleurs, concernant la relation existant entre mode de représentation et conception de l'objet étudié, il est à noter que la représentation en ligne est la plus utilisée par les interactionnistes. Cette représentation décrit / transcrit linéairement

[50] Citation originale : « An important error to avoid is the omission of certain signals from the measurement because they were prematurely considered to be meaningless. »

information verbale et non-verbale, en allant à la ligne à chaque changement de tour de parole. Elle met ainsi en avant l'aspect séquentiel de l'interaction (Ochs, 1979) en tant que système d'alternance de tours de parole. Par contre, elle limite l'analyse de l'interaction d'un point de vue multimodal (qui est quant à elle mise en valeur par les représentations utilisées dans les éditeurs d'annotations multimodales de type ANVIL ou ELAN –*cf.* partie IV.2. de ce chapitre-).

Cette tradition, qui transcrit les données, s'oppose à la tradition expérimentale. Cette dernière cherche en effet à coder, c'est-à-dire à tout noter de manière exhaustive, sans interpréter et sans décider *a priori* de ce qui est pertinent. Afin de rendre cette tâche possible, les recherches issues de cette tradition réduisent l'objet d'étude en utilisant d'autres moyens que celui d'estimer la pertinence *a priori* des éléments.

### II.3.3. Approche par la fonction des comportements

Dans certaines recherches (*e.g.* Cosnier, 1977, ou Kendon, 2004), l'approche utilisée pour étudier les comportements concerne leurs fonctions.

Par exemple Kendon (2004) se focalise sur la fonction communicative des gestes. Il utilise pour cela un modèle de type sémiotique, c'est-à-dire s'attachant à la valeur de signe des éléments considérés.

Il s'intéresse ainsi uniquement aux gestes en tant qu'« action visible utilisée comme énoncé ou partie d'énoncé »[51], un énoncé étant « tout ensemble d'actions considéré par les autres comme une tentative par l'"acteur" de "donner" de l'information de n'importe quelle sorte. » [52] (Kendon, 2004, p.7).

Kendon décrit donc uniquement les signaux significatifs pour la parole, et le centre d'intérêt de ses recherches réside dans la relation existant entre gestes et langage. i1 a ainsi effectué des micro-analyses de la coordination entre mouvements du corps et langage produit (Kendon, 1972).

### II.3.4. Approche par la signification des comportements

Dans d'autres recherches (*e.g.* Calbris & Porcher (1989) ou McNeill (1992)), les comportements non verbaux sont étudiés à travers leur sens.

McNeill s'intéresse principalement au lien existant entre geste et pensée. Il voit le geste comme découlant d'une « unité d'idée » (*cf.* Chapitre 2 I.2.1.).

[51] Citation originale : « visible action when it is used as an utterance (or a part of an utterance) »

[52] Citation originale : « any ensemble of actions that counts for others as an attempt by the actor to 'give' information of some sort »

> « [L]es gestes spontanés, idiosyncrasiques, imaginaires, sont des représentations de l'image mentale du locuteur. »[53] (McNeill, cité par Kendon, 2004, p.82)

Pour classer les gestes, il les interprète en termes de mode de représentation (*e.g.* iconique, emblématique, etc.).

D'autre part, McNeill (1972) est connu pour sa morphologie temporelle des gestes, largement reprise dans les modélisations du domaine de l'*Affective Computing*. Il décompose le mouvement en différentes phases successives, le « temps fort », l'apex, du geste étant appelé « *stroke* », et les autres phases étant organisées autour de lui. L'inconvénient de ce modèle est de ne pas rendre compte de la dynamique du mouvement à proprement parler, avec ses phases d'accélération et de décélération. La dynamique peut être analysée de manière pertinente à un niveau à la fois plus global et plus subtil qu'une succession de phases.

Quant à nous, nous cherchons à tout décrire exhaustivement et sans interpréter. Nous travaillons donc sur une situation d'interaction réduite au minimal, en gelant en particulier tout les comportements purement sociaux, afin de rendre la tâche possible.

### II.3.5. Décrire la valeur affective des expressions

La question globale est ici de savoir comment assigner un état au sujet en fonction de son comportement. Cela se ramène à deux sous-questions : sous quelle forme ? et de quelle manière ?

#### Sous quelle forme décrire un affect, et plus globalement l'état d'un sujet ?

Deux approches co-existent pour décrire les états des sujets expressifs, états qui sont en partie émotionnels et en partie cognitifs (Cowie, 2009) : la première approche, dite « catégorielle », utilise des catégories du langage courant pour décrire ces états, tandis que la seconde consiste à utiliser des dimensions.

Quelle que soit l'approche utilisée, il est nécessaire de considérer les changements graduels des états au cours du temps, ainsi que le phénomène de *mixed emotions* (littéralement « émotions mélangés »). Ainsi, attacher un label / une valeur à un échantillon pose à la fois des problèmes de procédure et de validation (Cowie, 2009) : les indices qui le permettent sont susceptibles d'être distribués à la fois au cours du temps et à travers différentes modalités. De plus, la décision finale du label peut fortement dépendre du contexte (*cf.* Chapitre 2 ).

Dans l'approche catégorielle, les expressions faciales correspondent à des émotions spécifiques, le plus souvent les catégories des émotions de bases (voir entre autres

[53] Citation originale : « [S]pontaneous, idio-syncratic, imaginistic gestures are representations of the mental images of the speaker. »

Ekman -Chapitre 1 IV.-). De l'autre côté, de nombreuses études adoptent une perspective dimensionnelle (*e.g.* Schlosberg (1954), Plutchik (1980) ou Woodworth (1938)). Dans cette approche, « les phénomènes émotionnels peuvent se décrire et s'expliquer en faisant appel à un ensemble de dimensions élémentaires qui se combinent pour produire n'importe quel état émotionnel » (Kaiser, Wehrle, & Schenkel, 2009). Wundt (1874) a proposé un système tridimensionnel pour caractériser les expressions : « agrément *vs.* désagrément », « excitation *vs.* dépression » et « tension *vs*. relaxation ». Les études ultérieures ont par la suite systématiquement confirmé les deux premières, respectivement la « valence » et l'« activation ».

Toutefois, même si les émotions peuvent être projetées dans un espace à deux dimensions ou plus, Sander & Scherer (2009, p.36) font remarquer qu'« il s'agit toujours d'une simplification. La quantité de labels verbaux pour décrire les émotions […] indique qu'une différenciation beaucoup plus subtile est possible. L'expérience subjective étant souvent limitée à l'expérience consciente de l'émotion, il semblerait que l'expression verbale est ce qu'il y a de plus proche pour la définir ».

Plus récemment, une autre approche, fondée sur la théorie de l'*appraisal*, a vu le jour (Cowie, 2009). Elle cherche à déterminer l'expression d'émotion par la description des évaluations qui en sont sous-jacentes. Elle est ainsi fondée sur l'idée que les signaux transmis à travers l'expression faciale, reflètent les éléments de l'*appraisal* de manière plus directe que les catégories d'émotions holistiques (Wehrle, Kaiser, Schmidt, & Scherer, 2000). Toutefois, il n'existe que peu d'études empiriques travaillant avec cette approche.

Concernant le réseau européen Humaine (Schröder & Cowie, 2006 ; Campbell et al., 2006, partie 6.1), il fait le choix de combiner plusieurs approches pour décrire les expressions : il utilise les catégories verbales du langage courant, des descriptions dimensionnelles larges, et des descriptions fondées sur la théorie de l'*appraisal*.

**De quelle manière assigner les états aux comportements du sujet ?**

Cette question est délicate car elle est intrinsèquement reliée au problème de l'interprétation du comportement.

Le débat concernant la légitimité scientifique d'interpréter le comportement des sujets étudiés relève du fondement des différentes disciplines concernées. Pour notre part, nous cherchons à éviter à tout prix cette interprétation.

La plupart du temps, lorsqu'il y a interprétation du comportement, un degré d'accord inter-juges est utilisé pour la valider. Cette méthode peut être utilisée au cours de différentes étapes des recherches. Alors que certains utilisent ce degré d'accord pour valider l'interprétation donnée par leurs « expert annotateurs », d'autres l'utilisent

uniquement en phase avancée de leur recherche, par le biais d'expérimentations perceptives, les juges étant alors externes.

Quoi qu'il en soit, un certain nombre de recherches portent sur les formes signifiantes d'unités « non verbales » (*e.g.* Birdwhistell, 1968, fondateur de la kinésique, ou Ekman & Friesen, 1975). Leur objet implique nécessairement une interprétation du comportement à un moment ou à un autre.

Nous verrons dans la partie III. comment éviter cette interprétation, et nous nous contentons ici de citer à titre d'exemples les travaux sur le FACS (Ekman & Friesen, 1976) et de travaux en dérivant.

### Le FACS, l'intérêt de son objectivité, et les dérives de son interprétation

Ekman étudie en particulier les expressions faciales. Il en fait une description « visio-musculaire » en utilisant son système FACS (Ekman & Friesen, 1976). Ce système permet une description objective des expressions faciales, qui n'est pas toujours adaptée selon nous (*cf.* partie IV.4.). Cependant, cette description a ensuite évolué vers une recherche puis une modélisation d'expressions émotionnelles prototypiques non écologiques.

Ekman décompose les expressions faciales en unités d'action (*Action Unit, AU*), pouvant être additives en termes d'information, dans certains cas. Il sera intéressant de tester cette hypothèse d'additivité sur nos données naturelles, et d'en dégager les règles combinatoires le cas échéant.

Matsumoto utilise quant à lui le FACS d'Ekman pour étudier l'influence de la culture et autres groupes sociaux sur les expressions faciales. Un point commun avec nous réside dans la granularité, à la fois locale et globale, des éléments qui sont pour lui de trois types[54] :

- les « micro-expressions », qui « se manifestent lorsque les individus dissimulent ou refoulent une émotion » ;
- les « macro-expressions », qui « se manifestent spontanément lorsque les individus n'ont aucune raison de prendre le contrôle ou de modifier leur expression » ;
- les « expressions subtiles » selon le paramètre d'intensité (et non selon le temps d'apparition de l'expression), qui sont, selon Matsumoto, « particulièrement importantes à distinguer pour les personnes qui souhaitent améliorer leur capacité à détecter le mensonge ».

[54] Définition et description des applications qui suivent trouvées sur le site de Matsumoto, dans la partie décrivant les produits qu'il commercialise, à l'adresse http://www.davidmatsumoto.com/research.php, consulté le 14/09/2010.

Toutefois, nous observons par les définitions mêmes qu'il en donne, que ses études sont fondées, sans étude systématique, sur l'interprétation du comportement de l'individu observé. L'interprétation de la valeur des expressions faciales de ces trois catégories fait l'objet de deux applications qu'il commercialise (par Humintell) :

- « Microexpression Recognition » (MiX) supposé aider à identifier et comprendre les comportements non-verbaux dans les situations, et améliorer la capacité à reconnaître les signaux d'émotions dans de multiples modalités ;
- « Subtle Expression Recognition Training » (SubX), censé aider à détecter le mensonge par le biais des expressions subtiles.

Il est à noter que ces deux systèmes revendiquent d'être efficaces dans les situations quotidiennes, alors que leur méthode est fondée sur le FACS et les *display rules* d'Ekman. Or le modèle d'Ekman est fondé sur les « émotions de base » universelles, étudiées en utilisant des expressions prototypiques et actées. Aucune preuve de l'efficacité de ce modèle pour les expressions naturelles n'existe à notre connaissance.

Ces propos sont renforcés par les critiques formulées par Pascal Lardellier (2008) à l'égard de cette culture de l'interprétation en pleine expansion. Ainsi, dans son livre « Arrêter de décoder », il dénonce ce qu'il appelle les « gourous de la communication ». Ce livre est le résultat d'une grosse enquête menée sur plusieurs années, notamment sur le système sous-jacent à ces « pseudo-théories » et ses enjeux socio-économiques. Ce système fait de cette culture de l'interprétation un domaine en vogue, entretenu par les médias et les personnes impliqués.

En somme, comme le résume Kaiser, Wehrle, & Schenkel (2009, p.107) :

> « Étant donnés la multi-fonctionnalité du comportement facial et le fait qu'aucun des différents modèles théoriques actuels n'est capable de prédire pour chaque émotion donnée, si et de quelle manière elle serait exprimée par un individu, un minimum de deux conclusions s'imposent pour la recherche empirique :
> - l'utilisation d'un système de codage indépendant de suppositions *a priori* concernant la signification des expressions (émotionnelle ou communicative) et permettant une approche analytique[...] ;
> - le sens d'une expression faciale ne peut être interprété que si l'on prend en compte l'intégralité temporelle et contextuelle d'une situation (comme, par exemple, les messages parallèles et/ou successifs verbaux et non-verbaux dans d'autres canaux que le visage). »

En partant des différents enjeux méthodologiques concernant l'étude du comportement expressif, nous avons passé en revue différents modèles à titre d'exemples, qui nous paraissaient caractéristiques.

Ces différents modèles ont tous été largement discutés, et confrontés à des tests utilisateurs lorsqu'une application en découle. Ils ont même été mis en regard les uns avec les autres, ou en regard avec d'autres modèles des émotions (*e.g.* les études de Kaiser, Wehrle, & Schmidt, 1998, et de Scherer & Ellgring, 2007). Cette confrontation des modèles est facilitée par les réseaux d'excellence, tel le réseau Humaine qui a cherché à faire une synthèse de ces différents modèles lors de simulations sur agents virtuels. Tous ces modèles ont ainsi montré leur légitimité.

Cependant, il est à noter qu'ils sont presque toujours construits initialement autour de noyaux dédiés à la description d'événements véhiculant des informations linguistiques et pragmatiques. Cette approche leur donne globalement une ligne « sémiotique ». Cela signifie que leurs démarches sont initialement déductives, même si elles se confrontent ensuite à la réalité des données, puis associent par induction des morphologies d'expressions à des valeurs informatives.

En parallèle, les risques de ce type de méthodologie déductive ont été montrés par des laboratoire particuliers de linguistique comme le Laboratoire Parole et Langage à Aix en Provence (Guaïtella et al., 1998), ou encore par d'autres domaines, comme l'éthnométhodologie (Garfinkel, 1967) et l'éthologie (Tinbergen, 1963 ; Pléty, 1993). C'est pourquoi nous nous sommes inspirés des méthodes utilisées dans ces deux domaines, et avons choisi une méthodologie qui se situe entre l'empirisme et l'expérimentation. À la lumière de ces considérations théoriques, issues de la philosophie, et méthodologiques, il nous semble nécessaire de ne pas « poser » le modèle avant d'avoir observé les données.

Nous ne prétendons en aucun cas proposer un nouveau modèle, mais nous cherchons à voir si certains mouvements, gestes ou autres indices pertinents, pourraient compléter les modèles existants. Notre but est donc de proposer nos observations, puis d'essayer de les intégrer à ces modèles.

# III. Vers une méthodologie inspirée de l'éthologie

La problématique que nous adoptons peut être étudiée avec une méthodologie qui se situe entre empirisme et expérimentation (*cf.* partie I.3. de ce chapitre). En effet, les primitives des comportements que nous observons n'ont jamais vraiment été observées dans le détail, et il s'agit pour nous de soumettre à l'expérimentation, c'est-à-dire à la reproductibilité en production et perception, des exemplaires de ces primitives. La méthodologie suivie est donc la clé de l'objectivation de notre démarche.

En 1998, Guaïtella et collègues cherchaient à analyser dans la parole les liens sémiotiques existant entre gestualité et vocalité, en s'ancrant à la fois aux travaux des interactionnistes (plus spécialement les ethnométhodologistes et éthologues), et ceux des linguistes et phonéticiens. La méthodologie de ce groupe de recherche insiste sur l'importance que le corpus soit spontané et que les hypothèses ne soient pas figées. Ils cherchent ainsi à « être attentifs à tout ce qui peut émerger des données. » (Guaïtella et al., 1998). Bien qu'ils travaillent sur de la parole et plus précisément en sémiotique, leur approche méthodologique est proche de celles des ethnométhodologistes et éthologues.

L'ethnométhodologie est une discipline créée dans les années 60 par Garfinkel (1967), qui contredit la sociologie classique. En effet, elle « insiste sur la capacité des individus à s'appuyer sur une connaissance ordinaire pour agir et rendre compte de leur action. » (Molénat, 2008). Comme tout humain communicant, nous, scientifiques, utilisons naïvement notre objet d'étude. Cela implique qu'il est nécessaire de limiter notre analyse de la société en général, et des expressions faciales en particulier, à la manière dont elle s'accomplit en situation, sans préjuger de ce qui s'y joue.

C'est dans l'éthologie, discipline étudiant le comportement animal (dont humain), que nous avons trouvé une méthodologie appropriée à nos préoccupations méthodologiques et à notre objet d'étude. Elle permet en effet de « pratiquer l'observation et l'étude des phénomènes de communication ou d'interaction en satisfaisant aux exigences d'objectivité et de rigueur qui caractérisent toute démarche scientifique » (Cosnier & Bourgain, 1993, p15).

Qu'est-ce que l'éthologie ? Quels sont donc ses principes ? Et que peut nous apporter cette approche ?

## *III.1. L'éthologie et sa méthodologie, et leur adaptation à nos recherches*

### III.1.1. L'éthologie et sa démarche

L'éthologie est une approche scientifique du comportement dans une perspective biologique. Alors que l'éthologie et la psychologie comportementaliste, issue du béhaviorisme (psychologie Stimulus-Réponse), semblent se confondre dans un objectif commun d'étude du comportement, elles s'opposent pourtant par leurs origines, leurs objets et leurs méthodes. Tandis que le béhavioriste se sert de l'animal pour vérifier expérimentalement des hypothèses concernant le comportement humain, c'est le comportement propre à l'espèce qui importe en premier lieu à l'éthologue.

Pour un éthologue, il existe quatre manières d'aborder un problème, qui sont toutes nécessaires pour avoir une compréhension globale de ce dernier. Elles sont plus connues comme « les quatre questions de Tinbergen » (Tinbergen, 1963) :

- l'approche phylogénétique, ou psycho-évolutionniste ;
- l'approche ontogénétique, c'est-à-dire développementale ;
- l'approche proximale, c'est-à-dire l'étude des causes immédiates, des mécanismes ;
- l'approche ultimale, ou fonctionnaliste.

Nous aborderons dans notre étude les points de vue ultimal et phylogénétique, et dans une moindre mesure le point de vue proximal. Nous gardons à l'esprit l'approche ontogénétique comme perspective à plus long terme.

L'objet de l'éthologie est le comportement manifesté par l'animal dans son écologie naturelle, son comportement spontané sur le terrain, par opposition au comportement réactionnel (ou provoqué) étudié en laboratoire par les psychologues comportementalistes. Sa méthode très spécifique, tout en étant largement fondée sur l'observation, est également expérimentale. Toutefois, l'observation initiale, bien que n'étant pas toujours une fin en soi, reste une étape fondamentale de la méthode éthologique. Ce sont les animaux qui posent les problèmes aux chercheurs et non l'inverse (Cosnier, 1977).

L'éthologue commencera son étude par le *sit and watch*[55], « s'efforçant de capter le maximum d'informations avec le minimum d'idées préconçues ».

Ensuite, un travail préliminaire à toute recherche éthologique consiste à établir des *éthogrammes*, « précisant avec autant de minutie et d'objectivité qu'il est possible, le

[55] Littéralement « s'asseoir et observer »

répertoire des unités comportementales [de l'animal] puis leur combinaison séquentielle. » (Cosnier & Bourgain, 1993, p.10).

Les traitements mathématiques ne peuvent éventuellement avoir lieu qu'une fois que l'étape fondamentale d'observation-description (ou étape *naturaliste*) a été effectuée. L'observation peut également être complétée par une étape *expérimentale* qui consistera à « manipuler les signaux (méthode des leurres) ainsi que l'environnement écologique et social (élevage en milieux enrichis ou appauvris; modifications du biotope, de la composition du groupe, etc.). » (Cosnier & Bourgain, 1993, p.10).

Un schéma représentant la démarche éthologique est proposée Figure 11.

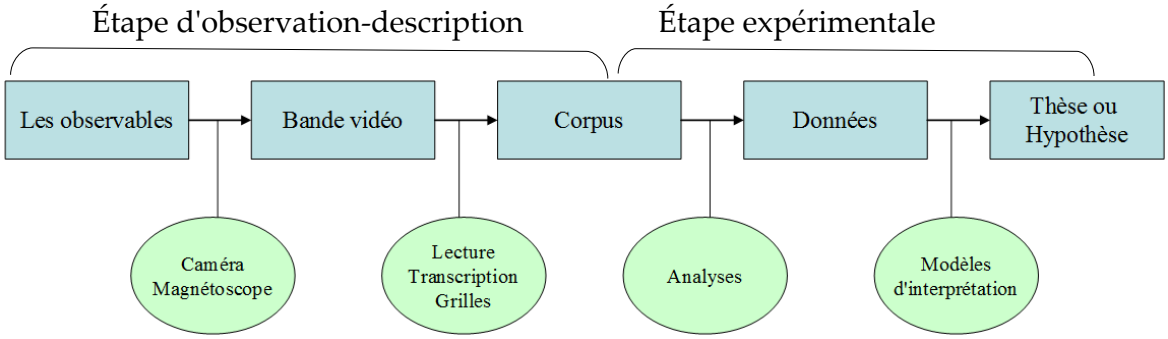



Figure 11: Démarche proposée par cette méthodologie éthologique (Pléty, 1993, p.22)

L'éthologie permet donc une « psychologie de l'interaction de l'individu en contexte » *(ibid)*.

La méthode éthologique a d'abord été transposée à l'humain dans l'étude des interactions infantiles et de la communication non-verbale, domaines par lesquels le rôle du langage parlé semblait pouvoir être contourné.

### III.1.2. Des objets d'attention : l'anthropomorphisme et l'empathie

L'anthropomorphisme est le fait de prêter des caractères, des spécificités humaines, à des formes, des individus non humains. Il s'agit d'une prédisposition naturelle, entretenue par la société (*e.g.* à travers les dessins animés, les bandes dessinées ou les contes, qui font parler, avoir des sentiments, des intentions, etc. divers animaux ou objets). L'anthropomorphisme se manifeste par la subjectivité (c'est-à-dire non neutre du point de vue de l'intention sous-jacente impliquée) des termes employés par l'humain pour tenter de décrire objectivement un comportement, notamment animal. Il pose le problème de la conscience attribuée à l'animal, qui apparaît lorsqu'un humain dit d'un animal qu'il est gentil, méchant, fier, etc. (Graff, 2008).

Il est à noter que l'anthropomorphisme apparaît également avec les technologies (*cf.* Chapitre 2 II.1.) : il s'agit du fondement de l'existence de l'interaction homme-machine

incarnée, physiquement (*e.g.* les robots « affectifs » de type Kismet - AILab-MIT - ou le chien Saiba - Sony -) ou virtuellement dans les agents virtuels.

Quoi qu'il en soit, un des grands principes éthologiques est de tenter de se dégager de l'anthropomorphisme et de plutôt parler en terme de stimulus-réponse pour décrire un comportement.

En travaillant sur l'humain le travers méthodologique à éviter lors de la description des comportements, est non plus celui de l'anthropomorphisme, mais celui de l'empathie.

Selon Cosnier (1994), l'empathie correspond à deux phénomènes différents :

- la « prise de rôle » (*role-taking*), qui serait la capacité de l'individu à se mettre à la place d'autrui, et à déduire ainsi mentalement ses pensées, ses sentiments et ses actions ;
- le « partage de perspective » (*perspective taking*), qui consisterait à imaginer ce que nous percevrions à la place d'autrui.

L'empathie est inhérente et indispensable à l'humain dans le sens où elle est à la base de toute activité sociale :

> « L'inférence empathique est "la voyance quotidienne" que chacun fait chaque fois qu'il tente d'inférer les pensées et sentiments d'autrui »[56] (Ickes, 2009, p.70)

Une personne empathique peut donc s'imaginer à la place de quelqu'un d'autre, et ainsi comprendre et prédire précisément les pensées, sentiments et actions de cet autre (Dymond, 1949).

Cette notion d'empathie est fondamentale en éthologie humaine, et plus globalement dans toute étude portant sur l'humain et son comportement, car dans ces domaines liés aux sciences humaines, un « expert » ne doit pas interpréter ce qu'il a pour tâche d'étiqueter. Pourtant, dans les corpus recueillis sur le vif (comme des entretiens journalistiques, etc.), il n'est pas toujours possible de séparer les niveaux d'information (lexique, expressivité vs. attitudes, intentions vs. émotions), ni de cacher le contexte à l'expert qui étiquette. Il est alors difficile d'éviter la subjectivité qui permet par exemple à un humain d'établir des prédictions. Par le fait même de sa nature humaine, un observateur d'une situation mettant en scène un ou plusieurs humains voit sa cognition et son affectivité modifiées par ce qu'il observe. Il faudrait que nous, observateurs humains qui devons étiqueter, arrivions à réfréner notre tendance naturelle à attribuer de l'affectivité aux autres humains.

[56] Citation originale : « Empathic inference is the 'everyday mind reading' that people do whenever they attempt to infer other people's thoughts and feelings ».

## *III.2. D'une méthodologie empirique et inductive...*

### III.2.1. L'observation

Le fondement de l'approche éthologique est le travail sur le terrain et non en laboratoire, afin d'avoir une observation naturaliste. Dans notre cas, le terrain « naturel » d'un soi-disant β-test[57] a pu être, paradoxalement, un laboratoire, puisque ce dernier est le lieu « naturel » d'un β-test de logiciel en fin de développement. Par suite, notre méthodologie s'attache à l'observation de comportements spontanés afin d'en établir des éthogrammes, ce qui correspond aux critères de Cosnier & Bourgain (1993) et de Pléty (1993).

Ainsi, notre travail porte sur le large corpus français d'expressions émotionnelles authentiques Sound Teacher / Ewiz (Aubergé, Audibert, & Rilliard, 2006), contrôlé par un paradigme d'induction de type magicien d'Oz, avec comme tâche prétexte un test de logiciel d'apprentissage des sons des langues du monde (*cf.* partie II.2.).

Nous sommes, selon l'éthologie, dans le cadre d'une « observation justifiée » avec outils visibles. C'est-à-dire qu'« une explication préalable et une négociation ont été réalisées avec les sujets », l'objectif exact de l'observation ayant été masqué, « soit en n'en laissant voir qu'un aspect secondaire, soit même en laissant croire à un tout autre motif que le motif véritable » (Pléty, 1993, p.27). Les avantages d'une telle méthode sont une moins grande inhibition et un comportement *a priori* plus naturel des sujets concernant le sujet véritable de l'observation.

De plus, les données recueillies sont comparables entre les sujets, puisque ces derniers se trouvent dans la même situation et utilisent presque toujours le même matériel sonore.

Finalement le paradigme du Magicien d'Oz adopté pour la constitution du corpus permet à la fois une situation naturelle conforme à l'approche éthologique, l'enregistrement en chambre sourde d'un corpus d'expressions émotionnelles au contenu contrôlé, et la présence visible d'une caméra et du matériel d'enregistrement physiologique.

En ce qui concerne le choix des sujets, il apparaît nécessaire de réfléchir aux limites de la population observée, et à ses caractéristiques propres, afin de ne pas en tirer des conclusions qui dépasseraient le cadre de l'observation effectuée. L'approche éthologique recommande de situer cette population dans un ensemble plus vaste pour

[57] Un β-test de logiciel est le dernier test d'un logiciel avant sa commercialisation, de manière à tout vérifier en situation et à éliminer les derniers bugs.

pouvoir en tirer des lois. Quant à nous, nous devrons garder à l'esprit cette remarque lorsque nous entamerons la mise en parallèle des données récupérées.

Par ailleurs, même si nous prenons les précautions indiquées pour le recueil des observables, l'objectivité n'est pas forcément acquise. Comme nous l'avons vu avec la notion d'empathie (partie III.1.2.), une observation objective est fonction de la lecture que le chercheur va en faire.

### III.2.2. Une réponse au problème de l'objectivité de l'étiquetage

La tâche d'étiquetage de notre corpus vidéo est particulièrement importante, dans la mesure où elle est le fondement de notre méthodologie empirique. La plupart des études sur les émotions menées jusqu'à présent en psychologie sociale, psychologie cognitive et en informatique affective, se sont fondées sur l'introspection ou l'interprétation intuitive des chercheurs (ou philosophes). Cette tâche d'étiquetage peut être décomposée en deux étapes, qui sont l'étiquetage proprement dit du comportement, et son interprétation ou son évaluation (ainsi que de son contexte). C'est cette deuxième étape qui pose un problème de subjectivité, bien qu'elle soit souvent conditionnée par la manière d'annoter / d'étiqueter à l'origine, donc par la première étape. Cette problématique découle du phénomène d'empathie.

Notre objectif, comme celui des éthologues, est d'être le plus objectif possible au niveau de l'étiquetage. Pour cela, nous adopterons les principes, cohérents avec ceux de l'éthologie, décrits dans la partie III.1..

L'utilisation de tels principes fera alors de notre transcription un éthogramme, qui consiste « en un répertoire d'actes et de postures observés et définis de façon précise par l'expérimentateur ; la grille d'observation est construite d'après cet éthogramme et permet de quantifier la fréquence des comportements sur une période de temps donnée avec, éventuellement, leurs durées et enchaînements. Chaque intitulé est défini selon des critères de direction, de sens, de localisation, de distance, d'intensité ou d'amplitude. »[58]

Certains systèmes d'étiquetage (proprement dit) du comportement utilisent des critères d'actions musculaires (*e.g.* le FACS d'Ekman & Friesen, 1976) ou d'articulation (*e.g.* Kendon (2004), pour les gestes). Malgré tout, le problème de l'objectivation pourtant indispensable lors de l'étiquetage / annotation des vidéos reste présent et rend la problématique de l'expertise mal posée dans notre domaine. C'est pourquoi la plupart des travaux d'étiquetage / annotation mettent en œuvre soit des tests de cohérence entre plusieurs experts, soit une vérification de l' étiquetage / annotation par des mesures perceptives effectuées par un panel générique d'auditeurs naïfs

[58] Définition trouvée sur internet à l'adresse http://gvaudan.ifrance.com/cptparen.htm, consulté le 17/02/2006

(Martin et al., 2006). Mais ces contrôles sont difficiles à mener systématiquement et rapidement sur de grands corpus (Kaiser, Wehrle, & Schenkel, 2009, p.83-85). Il nous faut par conséquent trouver une méthodologie afin d'être « experts » dans le domaine des émotions, au sens d'observateurs objectifs.

C'est pourquoi en plus de l'attention portée à la manière d'étiqueter les données (par les « experts »), nous utiliserons les auto-annotations effectuées par les sujets eux-mêmes à la suite de l'enregistrement comme « monnaie d'échange » lors de la phase d'expérimentation (*cf.* partie III.3.1.). Cela nous permettra ainsi d'avoir une autre source d'information que notre « expertise » sur les états exprimés par les sujets, et par conséquent de valider en quelque sorte l'étiquetage effectué.

Concernant l'étiquetage proprement dit, le scénario étant dans notre cas connu, et les états mentaux / affectifs étant auto-annotés par les sujets eux-mêmes, la subjectivité de l'étiquetage par un « expert » humain aurait pu être d'autant plus grande. C'est pourquoi nous avons fait en sorte que l'« expert » effectue l'étiquetage du corpus sans connaissance *a priori* de cette auto-annotation.

### III.2.3. Principe d'étiquetage

Nous n'avons pas utilisé une méthode d'étiquetage des expressions du sujet fondée sur une mesure automatique de l'image ou du signal vocal, mais une méthode restreignant le rôle de l'expert humain à la détection d'icônes gestuelles ou vocales minimales, sans interprétation de contenu informatif. L'expert étiquette chaque événement en utilisant ses compétences d'humain pour déterminer les moments de début et de fin de l'icône, de manière à ce que la granularité de l'étiquetage corresponde à un événement susceptible, d'un point de vue perceptif naïf d'humain, d'être pertinent en terme d'information.

Ainsi, si nous considérons les différentes manières d'étiqueter les formes, nous observons une gradation dans les contraintes appliquées aux observations, allant de l'analyse physique de mouvement (le traitement automatique) à l'analyse psycho-cognitive du mouvement qui va permettre de donner un sens à la forme (figure 12).

Quant à notre étiquetage, il se situe entre les deux : il est une sorte d'analyse « psycho-visuelle » et « psycho-acoustique » des formes. Il ajoute au traitement basique des signaux, un filtre cognitif de bas niveau, tel que celui qui nous fait percevoir des illusions ou celui de la Gestalt, non modélisé en traitement automatique à ce jour. Cette manière « psycho-visuelle / auditive » d'identifier les formes, liée à nos compétences d'humain, va ainsi permettre de quantifier cognitivement, et non physiquement, le mouvement, sans pour autant l'interpréter.

Le but de « l'expert » est donc d'utiliser ses compétences d'humain communicant pour sa tâche d'étiquetage, tout en minimisant sa compétence d'interprétation.

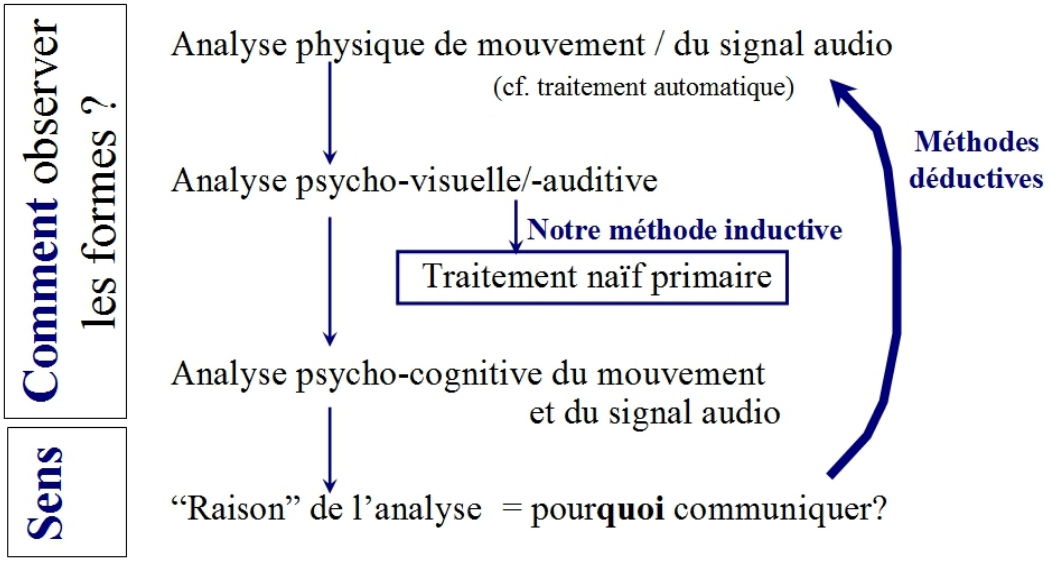



Figure 12: Schéma représentant les différentes manière d'observer les formes

Dans notre corpus, de nombreuses expressions vocales et / ou gestuelles apparaissent en parallèle, et notamment en dehors du tour de parole des sujets. Notre but n'est pas seulement de rechercher du code, c'est-à-dire des signaux communicatifs, mais aussi des indices des états mentaux et affectifs du sujet, qui sont toutefois toujours des événements significatifs et informatifs. Par exemple le bégaiement d'une personne est un repère de son comportement s'il est plus ou moins fort selon le niveau de stress. Dans ces cas là, ce que nous recherchons n'est pas en soi un encodage d'expression, mais seulement un indice de comportement, éventuellement idiosynchrasique.

En somme, nous nous situons entre :

- le modèle morphologique du geste de McNeill (1992), qui s'attache comme nous à la dynamique (mais en tant que succession de phases) et au lien entre gestualité et pensées. Cependant, son approche diffère de la nôtre par son objet d'étude (les gestes seuls) et par le type de classement utilisé, fondé sur l'interprétation des données (selon leur mode de représentation) ;

- le modèle de type sémiotique de Kendon, qui se focalise sur le lien entre gestes et langage, et effectue des micro-analyses sans influence de l'interprétation (en l'occurrence de la coordination entre mouvements du corps et parole). Néanmoins, son approche reste sémiotique avant tout et décrit uniquement les indices significatifs pour la parole, le geste étant déterminé par sa fonction communicative (« manifest deliberate expressiveness », Kendon, 2004, p.15) ;

- la description « visio-musculaire » d'Ekman & Friesen (1976), qui permet une description objective des expressions faciales. Malgré tout, elle ne nous semble pas toujours adaptée (*cf.* partie II.), et a dérivé vers une recherche puis une modélisation d'expressions prototypiques non écologiques.

Nous cherchons plutôt à établir un « modèle explicatif » plus global que ceux-là, qui a pour objectif de « cloner » certains comportements humains avec leurs cohérences globales et locales. Cela est rendu possible par notre démarche méthodologique qui ne présuppose pas d'axiomes au modèle. Cette démarche nous permet de nous dégager des *a priori* sur l'organisation temporelle, spatiale (dont morphologique), modale, ou de seuil d'amplitude du comportement.

Plus précisément, nous ne cherchons pas les fonctions des différents mouvements, mais souhaitons, à partir des événements observés dans notre corpus, dégager des icônes gestuelles ou vocales, dont il s'agira de vérifier la pertinence par rapport à l'évolution de l'état cognitif et affectif du locuteur, avant de déterminer leurs paramètres d'organisation temporelle et multimodale pertinents communicativement.

## *III.3. … à l'objectivation par l'expérimentation*

### III.3.1. Le choix de l'auto-annotation

Afin d'avoir une autre source d'information que notre « expertise » sur les états exprimés par les sujets, nous avons fait auto-annoter les vidéos par les sujets eux-mêmes, des études ayant montré qu'ils étaient les mieux à même d'évaluer les états dans lesquels ils se sont trouvés (Frijda, Kuipers, & ter Schure, 1989). En effet, la mémoire auto-biographique apparaît très fiable dans la mesure où les sujets sont sincères, et que son expression n'est pas contrainte (pas de choix fermé ou de limite d'espace pour s'exprimer) (Campbell et al., 2006).

Cette auto-annotation a été réalisée à la suite de l'enregistrement, sans contrainte en terme de contenu, ni de modalité. Les sujets avaient ainsi assez d'espace pour écrire ou pour dessiner s'il le souhaitait. Elle est pour nous un repère flou et étendu sur les états dans lesquels les sujets se souviennent avoir été. Il est à noter que les labels utilisés par les sujets sont des descriptions d'états complexes mélangés et même souvent quantifiés (*e.g.* « un peu stressée, perplexe, mais… »). Ils contiennent ainsi des références liées à plusieurs états émotionnels, états mentaux, humeurs, attitudes ou encore intentions. Ils nous servent (dans les cas où ils sont relativement simples) de monnaie d'échange lors des évaluations perceptives, par des juges naïfs externes, d'événements expressifs relevés (Chapitre 4 II.). Afin d'être fidèle à ce que les sujets ont dit ressentir, nous avons choisi de respecter les termes / expressions utilisés par les

sujets lors de ces évaluations, même dans le cas d'états complexes mélangés. En effet, rien ne nous dit que l'expression d'un état complexe soit la somme des expressions des états simples qui sont censés la composer. Nous n'avons donc ni découpé les auto-annotations en états plus simples, ni simplifié leur formulation en des termes plus génériques. Pourtant, cela a pu apporter des problèmes de signification des labels d'auto-annotation : cette signification n'était pas toujours partagée entre juges naïfs (ni entre le sujet et les juges naïfs).

Choisir l'auto-annotation naïve nous permet de nous échapper des pré-supposés théoriques difficilement évitables lors de la description de valeurs d'émotions, états mentaux, etc. (partie II.3.5.). Il s'agit ainsi de la première étape d'une démarche à la limite de l'« analyse-resynthèse » du *FoT*.

### III.3.2. Le rôle de l'expérimentation

Dans notre étude, l'expérimentation a pour objectif de valider les expressions que nous avons relevées. Il s'agit de vérifier si des icônes qui nous paraissent significatives le sont effectivement perceptivement, ou encore de savoir ce qui est emblématique dans nos données. Cela revient à laisser aux juges le soin de décider ce qui est significatif parmi les stimuli qui leur ont été proposés.

En effet, notre étiquetage naïf, issu d'un traitement psycho-visuel, est supposé avoir découpé de la matière pertinente à travers les icônes gestuelles. À ce niveau interviennent les auto-annotations en tant que monnaie d'échange entre sujet, envoyant de l'information de *FoT*, et juge recevant de l'information. L'expérimentation consiste ici à tester les icônes en perception auprès de juges naïfs. Si l'évaluation perceptive des icônes est cohérente entre les juges, la pertinence des icônes est donc validée et elles peuvent alors être considérées comme éléments à valeur communicative. À l'inverse, lorsqu'une icône obtient un taux de reconnaissance faible, cela signifie que ce morceau là en particulier, lorsqu'il est donné à percevoir, ne transmet pas assez d'information aux juges (ou pas de la bonne manière) pour qu'ils puissent évaluer sa valeur communicative.

Nous cherchons ainsi à établir des primitives communicatives, puisqu'à l'origine, l'objet de nos recherches (les icônes) n'est pas connu. Il est donc nécessaire de déterminer cet objet avant toute chose. Il s'agit d'une des grandes difficultés rencontrées lors de recherches de ce type.

Cette phase expérimentale ainsi que les résultats obtenus seront décrits en détail dans le chapitre 4.

### III.3.3. La nécessité d'une évaluation via les technologies

Bien que nous soyons intéressés par la compréhension du fonctionnement de l'humain, il est important de rester en contact avec les industriels, puisque l'application première des modèles réside dans les technologies (*cf.* Chapitre 2 II.1.)

Dans un travail à plus long terme, un autre type d'expérimentation pourra être par conséquent l'évaluation du modèle construit (ou des modifications de modèles existants) par l'utilisateur lui-même, dans le cadre de technologies d'IHM.

Dans cette perspective, le domaine de la PEI, « Psychologie – Ergonomie des Interactions » a pour objet d'étude les phénomènes d'interactions entre les usagers et les systèmes.

> « Pour une tâche donnée et pour une fonction donnée (logicielle et/ou matérielle) la PEI élabore des expériences qui permettent d'évaluer l'influence de la présence (ou de l'absence) de cette fonction sur la relation usager-système ». (Sansonnet, 2006, diapo 17)

Cinq critères principaux sont alors retenus, précisés dans la Erreur : source de la référence non trouvée : deux sont objectifs (l'efficacité et l'utilisabilité) ; trois sont subjectifs et donc intrinsèquement liés aux utilisateurs (la convivialité, la crédibilité et la confiance).

| | | |
|---|---|---|
| Objectif | **Efficacité** (efficiency) | mesure de la performance effective du couple usager- agent dans l'accomplissement de la tâche. |
| | **Utilisabilité** (usability) | facilité et capacité effective qu'a l'usager à bien comprendre comment fonctionne le système et donc à bien le commander. |
| Subjectif | **Convivialité** (user-friendliness) | est le « sentiment » qu'a l'usager que le système est agréable à utiliser (attrait, engagement, esthétique, confort). |
| | **Crédibilité** (believability) | est le « sentiment » qu'a l'usager que le l'agent peut comprendre ses problèmes et qu'il peut l'aider. |
| | **Confiance** (trust) | est le « sentiment » qu'a l'usager que l'agent se comporte comme une entité amicale et coopérative. |

Figure 13: Les cinq critères utilisés par la PEI (Sansonnet, 2006)

Ainsi, des études comme celle de Lester et al. (1997) ou de Beun, De Vos, & Witteman (2003) ont testé ce qu'ils appellent le « Persona Effect ». Ils ont comparé les évaluations (selon les critères sus-cités) d'utilisateurs d'applications pédagogiques : pour les premiers, testant la valeur ajoutée de la présence d'un agent (*vs.* son absence) ; pour les seconds, testant la présence d'un agent anthropomorphique, vs. d'un agent *cartoon*.

De Rosis (2001, p.275-276) met également en avant cette phase d'évaluation, en l'orientant quant à elle vers des tâches de comparaison entre le naturel (le corpus de données), et les résultats du modèle implémenté. Pour cette évaluation, elle insiste sur la nécessité de « développer des méthodes adaptées, surtout si le but n'est pas seulement d'évaluer le caractère "convaincant" d'une interface affective, mais aussi d'évaluer si les facteurs affectifs modélisés influencent le comportement des utilisateurs, "comme si" ils étaient en train d'interagir avec un humain, un interlocuteur affectif. »[59].

Quant à nous, notre objectif global est d'avoir un modèle fonctionnel, au sens courant (c'est-à-dire qu'il fonctionne), mais également au sens technologique. Or il est nécessaire pour cela de tester les technologies d'IHM dans lesquelles est implémenté le modèle auprès des utilisateurs eux-mêmes, qui en sont les vrais usagers. Ce sont eux qui valideront le réalisme de la technologie produite, en la rejetant ou en continuant à l'utiliser, et valideront ainsi le modèle sur lequel elle est fondée. Dès lors, ce seront les industriels qui nous donneront la définition de la réussite ou de l'échec d'une interaction.

[59] Citation originale : « This last point [evaluate] would require developing ad hoc methods, especially if the goal is to evaluate not only the "convincingness" of an affective interface but also whether affective factors do influence the Users behavior "as if" they were interacting with a human, affective interlocutor. »

# IV. L'étiquetage du corpus

Selon notre méthodologie empirique inspirée de l'éthologie, nous avons donc étiqueté les signaux audio-visuels du corpus SoundTeacher / E-Wiz, hors des tours de parole des sujets, sous l'éditeur d'annotation ANVIL, de manière à segmenter les mouvements en icônes gestuelles supposées minimales et fondées sur la variation des formes, sans interprétation.

## *IV.1. Le premier étiquetage linéaire*

Dans un premier temps, nous avons effectué un premier étiquetage linéaire sous le logiciel de traitement de texte Microsoft Word. Cette étape correspond en quelque sorte au « sit and watch » de l'éthologie. Nous avons décrit les vidéos le plus précisément possible, en suivant la démarche décrite précédemment et donc en attribuant à chaque « événement gestuel / facial » une « icône », accompagnée de plusieurs paramètres : sa description précise, son intensité, sa durée et sa position temporelle sur la vidéo.

Par ailleurs, nous avons superposé temporellement les événements faciaux et gestuels avec événements vocaux et verbaux, mais en gardant à l'esprit que le temps de production gestuelle est différent du temps de production verbale.

Nous avons ensuite complété cet étiquetage linéaire par l'étiquetage des quatre phases du scénario, les copies des écrans visualisés par les sujets et l'auto-annotation du sujet. Insérer ces informations au fil de l'étiquetage nous a aidés à corréler les données objectives liées au scénario, les données mentales/affectives fournies par les sujets et l'étiquetage de la gestualité du corps et de la face des sujets. Un extrait du premier étiquetage linéaire du sujet F_S se trouve Annexe 1.

Cet étiquetage linéaire a été effectué en collaboration étroite avec Fanny Loyau. Son analyse en terme de recensement d'icônes et de corrélation entre icônes a fait l'objet d'une partie de sa thèse (Loyau, 2007).

Cette première phase nous a permis :

- de recenser les différents observables de nos données, les différents types de comportements produits par les sujets ;
- de déterminer des icônes gestuelles primitives ;
- d'émettre des hypothèses quant aux paramètres et éléments susceptibles d'être pertinents d'un état mental/ affectif particulier du sujet, et donc nécessaire d'étiqueter ;

- de préparer le travail de formalisation, nécessaire à la création d'une grille d'annotation, et d'avoir des intuitions quant à la granularité et la précision auxquelles il est nécessaire d'étiqueter tel ou tel élément.

## *IV.2. Le choix de l'éditeur d'annotation*

À la suite de cette première annotation linéaire, il nous a fallu choisir un éditeur d'annotation à la fois fonctionnel pour notre type de recherche et adapté à nos contraintes, parmi ceux qui existaient au moment du choix, début 2006.

Un tel éditeur permet de réaliser par la suite un maximum de traitements automatiquement (comptage, etc.), et surtout de pouvoir visualiser et extraire des informations concernant les corrélations entre les différents niveaux temporels, les différentes modalités, les différents types d'événements, et entre événements et signaux physiologiques.

Nos recherches portent sur des données audio-visuelles et il nous fallait pouvoir exporter les annotations sous forme textuelle, pouvant ensuite être importées dans un tableur afin de les analyser. Nos autres contraintes étaient :

- de pouvoir créer un nombre de champs d'annotation important ;
- de pouvoir utiliser la même grille d'annotation (« fichier de spécification » sous ANVIL) pour différents sujets / vidéos, et pouvoir générer et retoucher ses champs facilement et sans contraintes ;
- de pouvoir utiliser des vidéos au format .avi[60] ;
- que l'éditeur gére des vidéos relativement longues (jusqu'à une heure) et occupant un espace-mémoire important (jusqu'à un giga-octet)[61] ;
- d'être précis dans le temps (c'est-à-dire pouvoir étiqueter à la trame près) ;
- de pouvoir afficher la forme d'onde du son (.wav) du fichier audio, ce qui facilite l'étiquetage des prises de parole et autres événements audibles ;
- d'être utilisé par les autres membres de la communauté scientifique afin de faciliter des échanges ultérieurs, échanges scientifiques oraux, mais aussi de données annotés et de grilles d'annotation (la « portabilité » des données) ;
- d'avoir la possibilité d'importer ultérieurement et de synchroniser les signaux physiologiques enregistrées, ainsi que des fichiers d'annotation de type Text-Grid, générés lors d'analyses acoustiques sous le logiciel Praat. C'est ce critère qui a été

[60] Des annotateurs, tel TASX, imposent soit la conversion de la vidéo de .avi en .mpeg, ce qui nous ferait perdre en qualité par compression, soit le chargement du son en .wav avant celui de la vidéo, ce qui nous poserait un problème de synchronisation son / image par la suite

[61] Il s'agit du point faible d'ANVIL

décisif dans le choix final de l'éditeur, et nous a contraint à abandonner l'annotateur ELAN, pourtant adapté à notre tâche.

Un rapport de Workshop[62] présente une comparaison des différents outils d'annotation multimodale utilisés dans la recherche, en particulier en sciences humaines, et nous a aidé dans cette tâche.

Nous avons finalement décidé d'utiliser l'éditeur d'annotation Anvil 4.5[63], élaboré par Michael Kipp au « German Research Center for Artificial Intelligence » (DFKI) et programmé en JAVA (*cf.* l'interface graphique d'ANVIL Figure 14).

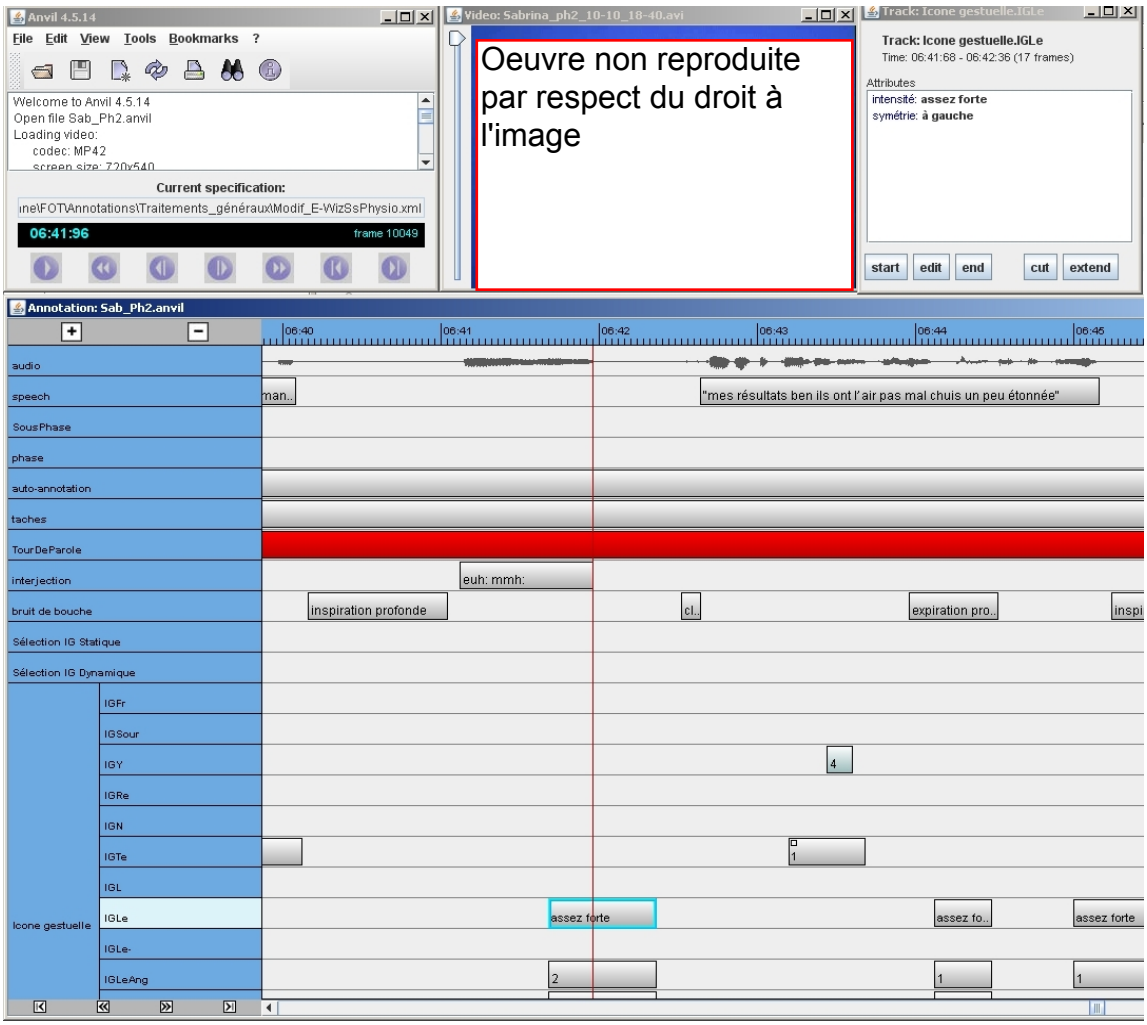



Figure 14: L'étiquetage du corpus, en utilisant l'éditeur d'annotations ANVIL

[62] Loehr, D (McLean), Duncan, S (Chicago), Rohlfing, K (Evanston) (2005) : "Comparison of Multimodal Annotation Tools -Workshop report". 2e Congrès de l'International Society for Gesture Studies (ISGS) INTERACTING BODIES - CORPS EN INTERACTION. École normale supérieure Lettres et Sciences humaines Lyon-France. 15-18 juin 2005

[63] Documentation sur ANVIL : Michael Kipp (2004), "Gesture Generation by Imitation - From Human Behavior to Computer Character Animation", Boca Raton, Florida: Dissertation.com; manuel utilisateur : http://www.dfki.de/~kipp/dissertation.html

Nous l'avons adopté en connaissant ses avantages et inconvénients : en dépit de la possibilité qu'il offre pour la création de champs et l'importation de données, d'une part il nécessite une phase préalable de programmation en xml pour établir un fichier de spécification, et d'autre part, il n'accepte pas les vidéos de plus de 20 minutes environ, pour une raison technique. Nous avons par conséquent été contraints de découper chacune des vidéos ainsi que leurs données associées selon les quatre grandes phases du scénario Sound Teacher, puis d'ajouter un champ contenant les informations concernant le fichier et son organisation temporelle.

Nous avons donc du définir les champs, leurs attributs, leurs relations, etc. dans le fichier de spécification (Annexe 2), en fonction des icônes gestuelles primitives établies préalablement. Ce fichier, bien que nécessitant du temps, a l'avantage d'être ensuite réutilisable pour tous les fichiers d'annotation.

## *IV.3. Étiquetage d'éléments du décours temporel*

Nous considérons ici des éléments d'ordre temporel liés au temps événementiel (c'est-à-dire liés aux événements dépendants du scénario), par opposition au temps communicatif (c'est-à-dire aux tours de parole du sujet). Nous reviendrons sur ces deux niveaux temporels dans la partie IV.1. du Chapitre 6.

Concernant le temps événementiel, les phases, au nombre de 4, et les sous-phases, au nombre de 44 à 46 (11 à 12 par phase), sont directement liées au scénario pré-établi (*cf.* partie II.2.). Elles ont chacune été étiquetée dans des *tracks* indépendants sous ANVIL. Concrètement, leurs frontières correspondent aux changements d'écran (consignes, exercice, ou résultats). Cela a permis de les étiqueter en s'alignant sur : les « bips » de synchronisation, correspondant aux changements d'écran à la suite d'une commande vocale « page suivante » émise par le sujet ; les dernières réponses des sujets pour chacune des séries de stimulus/réponse (puisqu'après cette dernière réponse, un écran indiquant les résultats de la série de réponses apparaît (pour plus de détails, *cf.* Audibert, 2004 et Aubergé, Audibert, & Rilliard, 2006). À chaque sous-phase peut correspondre un label d'auto-annotation (lié quant à lui au temps émotionnel) donné par le sujet lui-même (*cf.* partie III.2.2.). Nous avons associé à chacune des sous-phases un label correspondant à une « tâche récurrente » (que nous appellerons « tâche globale »). Trois états différents ont été distingués :

- réponse à une consigne donnée (notée « réponses » lors du traitement des données) ;
- résultats et commentaires du sujet sur la dernière série de réponses (« sous-résultats ») ;

- résultats globaux (avec éventuellement message d'alerte) concernant la phase complète (à visée inductive au niveau affectif), et commentaires associés du sujet (« warning »).

La Figure 15 schématise l'organisation globale des éléments du temps événementiel étiquetés dans la phase 3 (chaque phase est organisée à peu de chose près de la même manière).

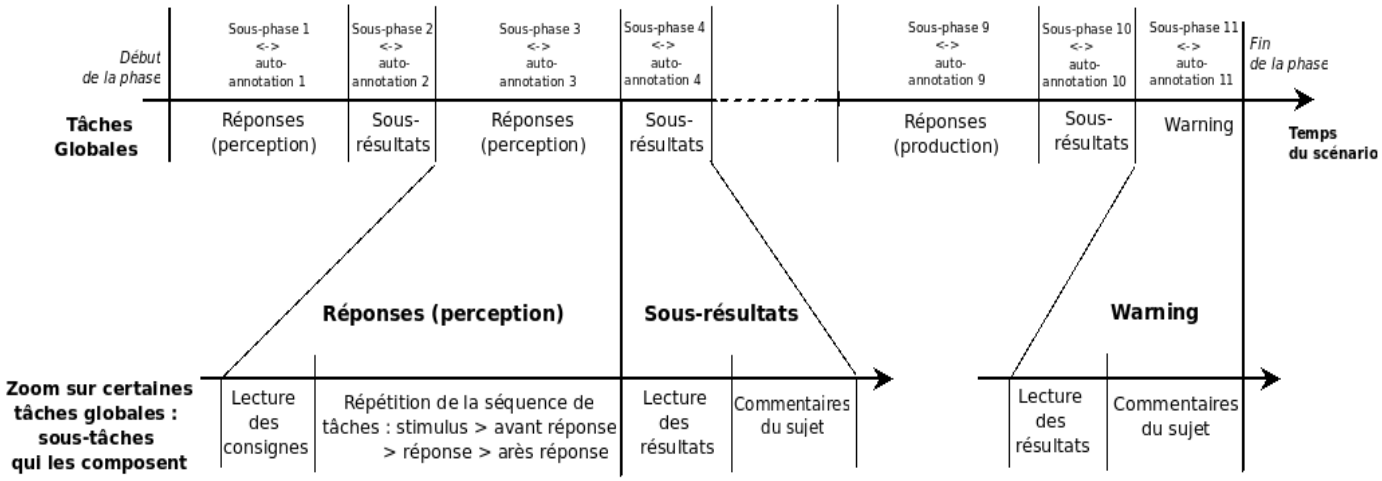



Figure 15: Schématisation d'une phase du scénario : exemple de la phase 3

D'un point de vue pragmatique, mais aussi cognitif, il était important de noter plus précisément la tâche (liée également au scénario) dans laquelle le sujet se trouve lorsqu'il produit une icône gestuelle ou un événement vocal.

Ainsi, un *track* particulier a été spécifié dans ANVIL, permettant à l'annotateur de sélectionner la tâche qui correspond à la plage sélectionnée, parmi celles présentes dans nos données de par le scénario. Les différentes tâches possibles sont :

- « pendant lecture »,
- « pendant page suivante »,
- « après page suivante »,
- « sur stimulus »,
- « pendant prononciation »,
- « entre prononciation »[64],
- « pendant commentaires »,
- « après commentaires »,
- « pendant réponse »,

[64] Lorsque le sujet répète (par erreur) le stimulus, la tâche a été étiquetée non pas « pendant prononciation », mais « pendant réponse »

- « avant réponse »,
- « après réponse ».

Certaines tâches sont spécifiques à une tâche globale : celles concernant les stimuli, les réponses et les prononciations se trouvent uniquement en tâche globale « réponses », et celles concernant les commentaires se trouvent uniquement dans les tâches globales « sous-résultats » et « warning ». À l'inverse, d'autres tâches sont communes aux trois types de tâche globale (celles concernant la lecture et la commande « page suivante »).

De même, la séquentialité des tâches peut être unique (*e.g.* le patron [« sur stimulus » - « avant réponse » - « pendant réponse » - « après réponse » ou « pendant lecture » de résultat-]) ; ou variable (*e.g.* « pendant page suivante » peut aussi bien apparaître après les tâches « pendant lecture » que « après commentaires » ou même « pendant commentaires »).

De plus, les tâches « entre prononciation », « après commentaires » et « avant réponse » impliquent forcément une prise de parole comme tâche suivante (respectivement « pendant prononciation », « pendant page suivante » et « pendant réponse »).

## *IV.4. La détermination des icônes gestuelles et leur codage*

Ce que nous nommons « sourire », « faux sourire » ou encore « grimace » dans la langue courante, en interprétant ce que nous percevons, pourrait probablement être des icônes gestuelles composées, qui serait issues d'une combinaison de ce que nous avons appelé icônes gestuelles primitives.

Du point de vue de leur nature, nos icônes gestuelles peuvent être globales (*e.g.* concerner les yeux) ou locales (*e.g.* un rictus au niveau des lèvres), et elles peuvent être l'étiquetage d'un mouvement comme d'une posture (*e.g.* « plisse les yeux » dont le temps de tenue est certainement un indice). Au niveau du codage, elles sont d'abord classées selon leur localisation, en utilisant un code de type « IG » (pour « Icônes Gestuelles »), suivi d'une abréviation pour la localisation (par ex. « Te » pour tête, « Sour » pour sourcil; ou « Le » pour lèvres) et enfin d'un chiffre correspondant au type de mouvement, dont la correspondance avec sa description est arbitraire. Par exemple nous avons « IGTe6 » pour « penche la tête sur le côté ». Cette notation permet d'être formel tout en facilitant l'apprentissage du codage pour les annotateurs.

Chaque type d'icône (IGXx) a donné lieu à un champ distinct (un *track*) dans ANVIL.

Nous avons ainsi décrit précisément chaque icône gestuelle, en précisant éventuellement sa direction, localisation, vitesse, symétrie, stabilité, répétition, intensité et/ou amplitude (Tableau 2). En effet, nous avons associé aux icônes des

variables (ou paramètres). Ces dernières sont soit constantes d'un type d'icône à l'autre (*e.g.* l'intensité ou la stabilité), soit spécifiques (*e.g.* l'ouverture de la bouche pour « IGLe »), soit communes à plusieurs icônes (comme la symétrie, pour « IGLe » et « IGSour »).

Tableau 2: liste des icônes primitives avec leurs variables associées

| Groupe ANVIL | IG | Pour… | Sous-catégories ou précision | N° | Type de mouvement | Variables |
|---|---|---|---|---|---|---|
| **Face** | IGSour | sourcil | | 1 | hausse sourcil(s) | Intensité, Symétrie, Vitesse, Répétition |
| | | | | 2 | fronce sourcil(s) | |
| | IGY | yeux | | 1 | écarquille 1 oeil/les yeux | Intensité, Symétrie, Vitesse, Répétition |
| | | | | 2 | plisse 1 oeil/les yeux | |
| | | | | 3 | ferme 1 oeil/les yeux | |
| | | | | 4 | cligne les yeux | |
| | | | | 5 | papillonne des yeux | |
| | IGRe | regard | direction | 1 | haut | Netteté (booléen), Intensité, Symétrie, Vitesse, Répétition |
| | | | | 2 | haut / gauche | |
| | | | | 3 | gauche | |
| | | | | 4 | bas / gauche | |
| | | | | 5 | bas | |
| | | | | 6 | bas / droite | |
| | | | | 7 | droite | |
| | | | | 8 | haut / droite | |
| | | | | 9 | sur la camera | |
| | IGN | nez | | 1 | fronce le nez | Intensité, Symétrie, Vitesse, Répétition |
| | | | | 2 | dilate les narines | |
| | | | | 3 | remonte l'aile/les ailes du nez | |
| | IGTe | tête | | 1 | hoche la tête | Intensité, Symétrie, Vitesse, Répétition, Précision avant/arrière |
| | | | | 2 | avance la tête | |
| | | | | 3 | recule la tête | |
| | | | | 4 | bouge la tête lateralement | |
| | | | | 5 | secoue la tête (type "non") | |
| | | | | 6 | penche la tête | |
| | | | | 7 | mouvement instable de la tête vers l'arriere | |
| | | | | 8 | grandit son cou | |
| | | | | 9 | tasse son cou, rentre le menton | |
| | IGL | langue et dents | | 1 | se passe la langue sur les lèvres | Intensité, Symétrie, Vitesse, Répétition |
| | | | | 2 | se passe les lèvres l'une sur l'autre | |
| | | | | 3 | se mord la lèvre inférieure | |
| | | | | 4 | se mord la lèvre supérieure | |
| | IGLe | lèvres | IGLe- | 1 | bouche en cul de poule | Intensité, Symétrie, Vitesse, Répétition |
| | | | | 2 | lèvres rentrees | |
| | | | | 3 | étire sa lèvre inférieure d'un côte | |
| | | | | 4 | descend sa lèvre inferieure | |
| | | | | 5 | avance sa lèvre inferieure | |
| | | | | 6 | remue les lèvres (lit en silence) | |
| | | | IGLeAng | 1 | angles des lèvres releves | |
| | | | | 2 | angles des lèvres abaisses | |
| | | | IGLeArr | 1 | lèvres protruses | |
| | | | | 2 | lèvres etirees | |
| | | | IGLeD (dents) | 1 | dents visibles | |
| | | | | 2 | dents uniquement inférieures visibles | |
| | IGJ | joue | | 1 | joue(s) gonflée(s) | Intensité, Symétrie, Vitesse, Répétition |
| | IGMe | menton | | 1 | menton froncé | Intensité, Symétrie, Vitesse, Répétition |
| | | | | 2 | mâchoire abaissée | |
| | | | | 3 | mâchoire sur le côté | |
| | | | | 4 | avance la mâchoire inferieure | |
| | IGB | bouche | | 1 | bouche ouverte | Intensité, Symétrie, Vitesse, Répétition |
| | | | | 2 | bouche fermee, lèvres pincees | |
| | | | | 3 | bouche fermee | |
| | | | | 4 | bouche entrouverte | |
| **Corps** | IGG | se gratte | | 1 | se gratte le dos | Main utilisée (gauche ou droite), Répétition, Vitesse |
| | | | | 2 | se gratte le front | |
| | | | | 3 | se gratte le nez | |
| | | | | 4 | se gratte la tête | |
| | | | | 5 | se gratte sous l'oeil gauche | |
| | | | | 6 | se gratte sous l'oeil droit | |
| | IGCO | cheveux derrière l'oreille ou main devant le front | | 1 | Cheveux derrière oreille gauche | Main utilisée (gauche ou droite), Répétition, Vitesse |
| | | | | 2 | Cheveux derrière oreille droite | |
| | | | | 3 | Cheveux derrière oreille gauche puis droite | |
| | | | | 4 | Cheveux derrière oreille droite puis gauche | |
| | | | | 5 | Main devant le front | |
| | IGBu | buste | | | | Répétition, Vitesse, Précision avant-arrière (avance ou recule le buste), Symétrie, Haussement (booléen) |
| | IGBG | autres mouvements | | Ep | bouge l' (les) épaule(s) | Symétrie, Répétition, Vitesse, Précision avant-arrière (avance, recule ou autre), haussement (booléen) |
| | | | | Br | bouge le (les) bras | |
| **Idio-synchrasique** | IGMeB | | sujet F_S | | bosse sur le menton | Intensité, Répétition, Vitesse |
| | IGF | | sujet F_S | | fossette | Intensité, Répétition, Vitesse, Symétrie |

Concernant l'amplitude du mouvement, elle est pour nous une variable paramétrée et relative à chacun des sujets qui nous sert seulement d'indication. De perception subjective, cette variable va voir ses valeurs ajustées au sujet intuitivement par l'« expert », après quelques minutes de vidéo étiquetées.

Nous nous attachons ainsi à avoir la perception psycho-visuelle décrite partie III.2., c'est-à-dire que nous ne prenons pas de décision quant à la pertinence d'un mouvement selon un certain seuil d'amplitude, ou en d'autres termes que nous étiquetons tout ce qui nous est visible (ou audible dans le cas des icônes vocales). Ce choix méthodologique s'oppose ainsi à celui de Kendon, pour lequel l'amplitude sert dans la détermination d'un seuil de signification du mouvement (Kendon, 2004).

Le FACS d'Ekman (Ekman & Friesen, 1976) pourrait être une bonne base en vue de formaliser une description fine des expressions du visage, puisque ce système de codage a été conçu d'une manière très précise et très méthodique, empêchant tout oubli d'un point de vue musculaire. Toutefois, nous ne l'avons pas utilisé, pas plus qu'un autre modèle pré-existant (*cf.* la comparaison des différents schémas de codage des expressions faciales effectuée par (Cohn & Ekman, 2005), pour deux grandes raisons : d'une part afin de ne pas être influencé *a priori* sur une décision sur la pertinence communicative des événements ; d'autre part afin de pouvoir adapter notre grille d'étiquetage à nos données, nos critères et nos besoins.

De plus, a posteriori, il semblerait en effet que le FACS ne soit pas directement adapté pour coder certains des mouvements que nous observons. Tout d'abord, il est conçu pour décrire des images d'un visage statique, et non des mouvements. De plus, certaines Unités d'Action (*AUs*) ne semblent pas pertinentes dans le sens où elles sont quasiment impossible à distinguer entre elles, ou sont trop précises pour être pertinentes dans la communication. De surcroît, certaines de nos icônes ne peuvent être décrites en utilisant ce système. Ainsi, nous avons effectué un groupement en classes d'icônes parfois différents du FACS : par exemple les icônes concernant les yeux (plissés, écarquillés, etc.) séparées de celles concernant le regard par rapport à l'écran. Nous avons également adapté notre description à notre objet d'étude, c'est-à-dire à une tâche d'interaction personne-machine, ce qui implique des sujets assis face à un écran. Nous avons ainsi traité ensemble les mouvements du cou et ceux de la tête puisqu'il n'était pas pertinent de les traiter séparément dans notre cas. Par ailleurs, nous avons parfois étiqueté plus finement que ce que le permet le FACS. Les raisons en sont :

- soit que les icônes ne peuvent être décrites par le FACS (*e.g.* « dents visibles » ou non, « se passe la langue sur les lèvres », « se passe les lèvres l'une sur l'autre », « se mord la lèvre supérieure »)(*cf.* exemples Figure 16) ;

- soit que l'icône est une particularité visible sur le sujet, conséquente à un mouvement générique particulier (*e.g.* la présence de fossettes) ;
- ou encore que l'icône est une résultantes physiques d'un état physiologique (*e.g.* « dilate les narines » ou « rose au joues »).

Nous tâcherons de démontrer en partie la pertinence communicative de la finesse de notre étiquetage, mais dans tous les cas, nous gageons que la naturalité de ces gestes passe par la finesse de notre description.


Figure 16: Deux exemple d'IGs a priori difficilement descriptibles en termes d'AUs
(à gauche, sujet T, label « angoissée »;
à droite sujet S, « Pas concentrée et envie de rigoler »)

Toutefois, il y a évidemment un grand nombre d'équivalences entre nos icônes d'une part et le FACS et autres modèles de gestualité d'autre part. Dans la mesure du possible, une correspondance entre nos icônes et unités d'actions (AU) du FACS et FAP's a été établie, étant donné l'objectif à moyen et long terme d'implémenter nos icônes sur des agents de type GRETA (Poggi, Pelachaud, De Rosis, Carofiglio, & De Carolis, 2005), dont le codage des mouvements dans la norme MPEG4, est fondé sur les travaux d'Ekman. Cette correspondance partielle se trouve en annexe p.296 (issue de Loyau, 2007, p.193). Une perspective à plus long terme sera de traduire intégralement nos icônes dans ces modèles, mais seulement après avoir trouvé ce qui a un effet communicatif. C'est uniquement en étiquetant aussi finement, que nous ne passerons pas à côté d'une variable qui pourrait être pertinente.

# V. Résumé

Étiqueter les formes des expressions et annoter leur valeur sont des tâches qui peuvent être guidées par une théorie, ou à l'inverse, par les données dans une approche inductive. Après avoir donné un aperçu des enjeux méthodologiques de ces tâches, et plus globalement des études du comportement expressif, nous avons présenté l'approche et la méthode éthologique, dont nous nous sommes inspirés pour construire notre propre méthodologie.

Nous faisons le pari d'une démarche d'abord inductive, avant de poser, d'en déduire, ou de tester des hypothèses sur les informations affectives. Nous tentons ainsi une observation de données écologiques, sans *a priori*, sans préjugé sur leurs valeurs affectives, sans modèle sur la morphologie des expressions, ni même sur leur possibles unités. L'utilisation d'une telle méthodologie permet l'objectivation de notre objet d'étude.

Nous travaillons sur le corpus d'interactions personne-machine d'expressions émotionnelles authentiques Sound Teacher / Ewiz (Aubergé, Audibert, & Rilliard, 2006). Nous avons veillé à rester des observateurs humains naïfs et à ne pas interpréter les comportements observés lors de son étiquetage. Ce dernier consiste ainsi pour des « experts », à « remarquer » et étiqueter finement et exhaustivement chaque indice, geste, mouvement, événement acoustique, en s'abstenant d'anticiper la moindre identification ou relation sémiotique.

Même si certaines recherches étudient déjà en détail des mouvements significatifs subtiles, en particulier en ce qui concerne la face (Ekman, Matsumoto, Kendon), la plupart de de ces modèles semblent *a priori* insuffisants pour décrire certains événements pouvant être rencontrés dans notre corpus.

Notre but n'est cependant surtout pas de nous opposer aux connaissances déjà très solides des domaines de la gestualité. Nous ne prétendons pas non plus entrer dans la morphologie complexe du geste et de sa simulation, à la manière de Martin et Pelachaud (*e.g.* 2006) suite aux grands modèles classiques de la gestualité.

Il s'agit au contraire, de transposer au final nos observations dans les modèles les plus résistants, après avoir validé leur pertinence et celle des paramètres étiquetés par des tests perceptifs et des analyses. Ainsi, nous espérons pointer sur des « détails » filtrés *a priori* et qui pourraient être pourtant très pertinents en IHM.

# CHAPITRE 4 : PERCEPTION DES MODALITÉS GESTUELLES / FACIALES

## I. La perception visuelle : des formes aux expressions faciales

La perception des formes et des objets, pourtant réalisée sans difficulté consciente par les observateurs humains, est en fait un processus très complexe. L'humain est en effet en général capable de reconnaître et identifier des objets, et en particulier des visages, sous différents points de vue et diverses luminosités, même s'ils sont en partie cachés. Modéliser cette capacité et l'intégrer aux machines fait partie des enjeux actuels.

### *I.1. Une perception globale*

Dans certaines expériences de biologie comparative, il a été montré que contrairement aux singes qui utilisent des bouts restreints d'image et ignorent une grande partie de l'*input* visuel, les humains récupèrent de l'information visuelle à partir de larges zones de l'image, et cela avec une grande robustesse (Nielsen, Logothetis, & Rainer, 2006). Ils perçoivent ainsi ce qui les entoure de manière globale. Pour un survol descriptif plus détaillé des différentes théories de la perception visuelle, se reporter par exemple à Streri (2001, p. 137 à 168).

#### I.1.1. Théorie de la Gestalt

Une théorie marquante portant sur l'organisation perceptive provient du mouvement gestaltique allemand (*Gestalt* signifiant « forme » en allemand), dans les années 20'. Cette « psychologie de la forme » s'intéresse à la manière dont le système humain organise cognitivement l'environnement. Elle cherche entre autres à comprendre par quels mécanismes l'individu parvient à détecter des invariants lorsqu'il est confronté à une multitude d'informations (*e.g.* reconnaître un ancien ami au milieu d'une foule, ou encore identifier des paroles humaines dans un brouhaha de la rue) (Weil-Barais, 2001, p.43). L'article fondateur de cette théorie, écrit par von Ehrenfels (1890), souligne l'idée qu'une forme est autre chose, ou quelque-chose de plus, que la somme de ses parties. Cette idée est étroitement lié au concept philosophique d'émergence, datant du XIXème siècle. L'émergence est une théorie « selon laquelle la combinaison d'unités d'un certain ordre réalise une entité d'ordre supérieur dont les propriétés sont entièrement nouvelles » (Foulquié, 1962)[65].

[65] Cité par Goldstein dans un article du site « Café Philo des phares » (http://www.cafe-philo-des-phares.info/index.php?option=com_content&task=view&id=172&Itemid=37) le 13/04/2008

Les gestaltistes, dont les trois plus grands noms sont Max Wertheimer (1912), Wolfgang Köhler (1929) et Kurt Koffka (1935), ont appliqué cette idée à la perception. Ils soulignent aussi qu'une partie d'un tout est perçue différemment lorsqu'elle est isolée ou incluse dans un autre tout, car elle tire des propriétés particulières de sa place et de sa fonction dans le tout (Streri, 2001, p.147). En somme, il est indispensable d'analyser la perception de tout événement en tenant compte du contexte global.

Une des grandes entreprises de cette théorie a été de décrire la perception que l'on aura des formes, suivant certaines lois structurales ou règles universelles, et qui s'imposent au sujet lorsqu'il perçoit. Toutefois, les lois Gestaltiques n'expliquent pas tous les phénomènes, en particulier lorsque les préoccupations scientifiques passent de la perception de la forme à celle des objets.

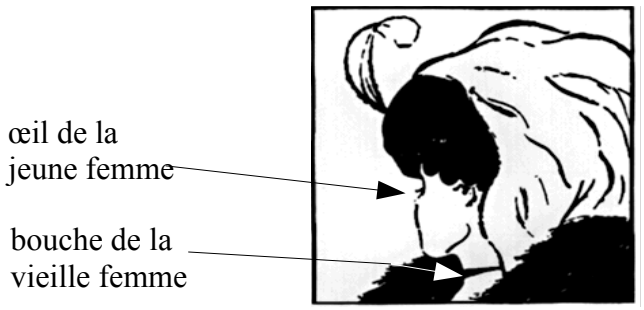



Figure 17: Figure ambiguë pouvant représenter une jeune femme ou une vieille femme

Par exemple lorsque la figure ambiguë (Figure 17) est donnée à percevoir, une des deux représentations apparaîtra spontanément (la jeune femme ou la vieille femme) et cela dépendra de l'humain qui perçoit. Toutefois, si nous focalisons notre attention sur certains détails, l'organisation spontanée initiale, et donc le processus d'analyse des figures, peuvent s'en trouver modifiés.

> « L'idée que la fixation d'une partie du champ visuel peut déterminer la perception d'une figure plutôt qu'une autre revient à donner à l'observateur un rôle dynamique dans l'organisation de son environnement, une capacité que lui refusait la théorie de la Gestalt. » (Streri, 2001, p.154).

De cette manière, la perception d'une scène aurait lieu en deux étapes, lors desquelles l'observateur aurait un rôle dynamique : d'abord la prise d'information à partir des parties de l'environnement, puis un regroupement de ces informations pour former la perception d'un ensemble cohérent. Pour des détails, voir les trois principales approches qui tentent de déterminer les unités d'analyse de notre système de perception visuelle : celles de Marr (1983), de Treisman (1988), et de Biederman (1987).

### I.1.2. Une perception multimodale

Dans ce chapitre, nous nous focalisons sur la modalité visuelle. Toutefois, il est nécessaire de garder à l'esprit que l'humain possède, au niveau de sa perception, une capacité à intégrer des *inputs* provenant de modalités variées :

> « Un aspect important du mécanisme perceptif de l'humain est sa capacité à intégrer des *inputs* de modalités sensorielles variées (*e.g.* la vision, l'audition, le touché, le goût). La manière dont nous percevons notre environnement est essentiellement multimodale, notre cerveau fusionnant les modalités pour construire un percept cohérent. »[66] (Swerts & Krahmer, 2006, p.1)

Les objets donnés à percevoir à l'humain sont multimodaux, la perception de ce dernier s'est donc adaptée : c'est ce qui fait que la manière qu'a l'humain de percevoir l'environnement est principalement multimodale (voir Chapitre 2 I. et Chapitre 5 I.).

## *I.2. Perception du visage*

Le visage est un élément particulier du corps humain, dans le sens où il est le moyen le plus fiable d'identifier un individu connu. De plus, le visage informe celui qui le perçoit sur notre identité (notre sexe, notre âge, etc., cet autre est un inconnu), mais également sur nos intentions, notre humeur, nos états mentaux, nos états émotionnels ou même notre personnalité. L'information que le visage porte est importante : il a été montré que les visages attirent plus l'attention que les objets (Bindemann, Burton, Langton, Schweinberger, & Doherty, 2007). De plus, un contrôle complet de notre manière d'observer n'est pas possible en présence d'une image de visage (*ibid*).

Au niveau du traitement du visage, en vue d'identifier les personnes, il existerait deux processus : le premier concerne l'encodage et la représentation en mémoire des visages non familiers, qui permet au second processus, la reconnaissance proprement dite d'un visage particulier, de se faire par la suite. Par ailleurs, il y aurait une modularité (au sens de Fodor (1983)) des traitements du visage : l'analyse de l'expression faciale, l'identification de la personne, ou encore la lecture labiale lorsque la personne parle (Campbell, Landis, & Regard, 1986) s'opèreraient chacune de manière indépendante.

Plus précisément pour l'identification des visages, l'identité d'une personne apparaît davantage spécifiée par la configuration des traits faciaux, que par les traits eux-mêmes. De plus, les différentes parties du visage ne sont pas d'égale importance : alors que « les traits externes des visages (contour, forme du visage et chevelure) sont

[66] Citation originale : « One important aspect of the human's perceptual mechanism is its capacity to integrate input from various sensory modalities (*e.g.* vision, audition, touch, taste). The way we perceive our environment is essentially multimodal in nature as our brain fuses modalities to produce a coherent percept. »

pertinents pour la reconnaissance de visage non familier, la reconnaissance de visages familiers repose davantage sur la perception de traits internes (les yeux étant plus importants que le nez) » (Streri, 2001, p.166-167).

## *I.3. Perception des expressions faciales émotionnelles*

Outre son importance pour reconnaître une personne, le visage, par le biais des expressions faciales, est un moyen privilégié pour transmettre de l'information à propos de ses états émotionnels, mais aussi de son humeur, son degré d'attention ou son attitude vis à vis de l'interaction ou plus généralement de la situation, ou encore de ses intentions.

> « Les expressions faciales reflètent les états de préparation à l'action, la volonté d'interagir, et peut aider un interactant à ajuster son comportement à l'état courant du *sender*. Mais les expressions faciales peuvent aussi être utilisées de manière consciente pour commenter les situations observées et les actions des autres personnes. »[67] (Grammer & Oberzaucher, 2006, p.3)

Malgré l'étendue des informations pouvant être transmises par les expressions faciales, la grande majorité des études porte sur les expressions émotionnelles. Parmi elles, plusieurs ont plus particulièrement porté sur la spatialisation de l'information sur le visage. Les premières ont été Ekman et Friesen (1976) qui avant de construire le FACS (*cf.* Chapitre 3 II.) ont recensé les modifications locales du visage qui apparaissent lors de la contraction musculaire qui accompagne un état émotionnel donné. Par la suite, quelques recherches ont cherché à préciser le rôle des parties supérieures et inférieures du visage dans la reconnaissance des différentes expressions émotionnelles (Bassili, 1979 ; Calder, Young, Keane, & Dean, 2000).

La mise en commun de leurs résultats sur les émotions de base (Baudouin, 2001) indique qu'alors que la colère et la peur sont mieux reconnues à partir de la partie supérieure du visage, la joie et le dégoût le sont mieux à partir de la partie inférieure, et la surprise est aussi bien reconnue dans les deux parties. Quant à la tristesse, les résultats sont différents selon les études (reconnaissance meilleure à partir de la partie supérieure du visage dans l'étude de Calder et al. (2000) *vs*. aussi bonne dans les deux parties du visage dans la recherche de Bassili (1979)).

Ces études concluent donc à un processus de reconnaissance des différentes expressions faciales émotionnelles déterminé par des modifications locales de certains traits faciaux, dont la description en vue de l'interprétation est de nature

[67] Citation originale : « Facial expressions mirror the states of action readiness, the willingness to interact, and can help an interacting person to adjust his behaviour to the sender's current state. But facial expressions can also be used consciously in order to comment on observed situations and other person's actions. »

componentielle. Par la suite, la question du rôle des différents composants et de leurs relations a été peu abordée, peut-être de par la difficulté de la tâche et l'équivocité des résultats déjà observés. Dans une évaluation de l'intensité émotionnelle de visages, pour lesquels ils faisaient varier un seul ou plusieurs traits d'un visage neutre, Wallbott & Riccibitti (1993) ont montré que les différentes émotions étaient exprimées à travers la variation d'une configuration de traits plutôt que celle d'un seul. À l'inverse, selon Ellison & Massaro (1997) il est possible de reconnaître une expression émotionnelle particulière à partir de la combinaison de variations locales. Toutefois, l'étendue de leurs résultats est restreinte puisqu'ils n'ont manipulé que deux traits : les sourcils et les coins de la bouche.

Les résultats de Calder, Young, Keane, & Dean (2000) renforcent quant à eux le rôle important des informations configurales dans la reconnaissance de l'expression faciale émotionnelle, en utilisant une technique de visages composites pour générer les stimuli : chaque moitié, supérieure ou inférieure, du visage d'une même personne, exprimait une émotion différente de l'autre ; les stimuli sont ensuite donnés à percevoir dans deux conditions différentes : les deux moitiés soit alignées, soit décalées (voir Figure 18). Ils ont ainsi montré que la reconnaissance de l'émotion exprimée sur chacune des moitiés du visage est perturbée lorsque les moitiés sont alignées, par rapport à la condition « décalé ». Les informations configurales semblent donc jouer un rôle prépondérant dans la reconnaissance de l'expression faciale émotionnelle.

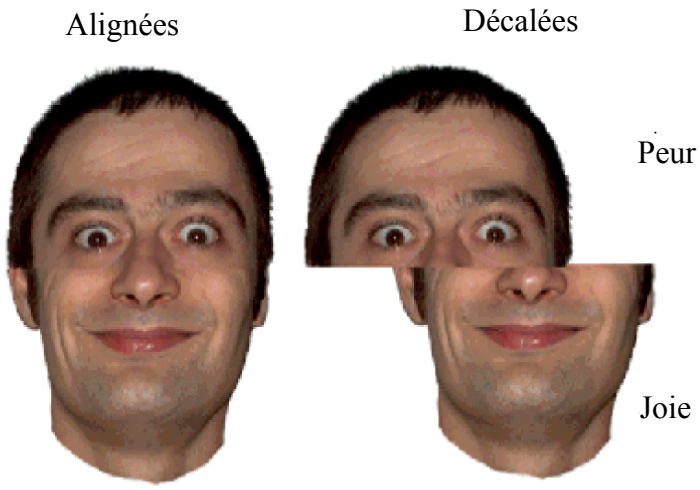



Figure 18: Exemple de stimuli créés par la technique de visages composites, telle qu'utilisée par Calder et al. (2000)

En parallèle, De Gelder & Vroomen (2000) ont reporté que l'information visuelle située dans la partie basse du visage (en particulier dédiée à transmettre de l'information linguistique) était moins importante que celle apportée par la région des

yeux pour la perception des émotions. L'originalité de cette étude réside dans la manière d'analyser les résultats. Elle a permis une interprétation différente, pouvant s'appliquer aux autres études également. Ils ont pris le problème non plus par les expressions faciales elles-mêmes, mais par la stratégie adoptée par l'observateur pour percevoir et interpréter l'émotion exprimée. Ainsi, ils ont montré que cette prédominance d'une partie du visage dans la perception des émotions était liée à la fois à un « effet locuteur » (*i.e.* la manière dont est produit l'expression) et à un « effet observateur » (*i.e.* les processus mis en jeu dans la perception de l'expression par l'observateur).

Il est dommage de constater que la plupart de ces études se sont focalisées sur l'expression des émotions de base, et de surcroît sur des expressions actées, et non spontanées.

Ce n'est que plus récemment, avec le regain d'intérêt pour la modélisation du comportement non verbal, que des études se sont intéressées à la configuration des expressions faciales, non plus émotionnelles, mais à fonction de focus verbal par les modalités visuelles : aux « facial markers of prominence in spoken utterrances ». Swerts et ses collègues (Swerts & Krahmer, 2006 ; 2008) ont cherché à étudier la fonction de différentes zones du visage : partie supérieure vs. inférieure du visage, et partie gauche *vs.* partie droite du visage.

En effet, concernant la répartition des informations de focus dans la première distinction de zones, de nombreuses recherches ont suggéré que les mouvements de sourcils pouvaient indiquer des focus sur certains mots au sein d'un énoncé (Ekman & Oster (1979) suivi par exemple de Cassell, Vilhjálmsson, & Bickmore (2001) ou Pelachaud, Badler, & Steedman (1996)). Plus globalement, les paramètres prosodiques tendraient à être localisés dans la partie supérieure du visage : des observateurs dans une tâche de prise de décision concernant l'intonation et l'accentuation passent plus de temps à regarder la partie supérieure du visage que dans une tâche de décision lexicale (Lansing & McConkie, 1999).

L'équipe de Swerts et Krahmer a donc mené cette étude du focus en cachant, pour chaque stimulus, une partie du visage du bord de l'écran jusqu'à une ligne verticale ou horizontale passant par le milieu du nez du locuteur. Elle a ainsi confirmé que la partie supérieure du visage apportait davantage d'éléments pour la détection du focus, que ceux apportés par le bas du visage. Elle a également montré que la partie gauche du visage était plus importante que la partie droite du visage. La comparaison des résultats de la perception des vidéos originales *vs.* des vidéos en miroir, a révélé qu'il s'agit à la fois d'un effet locuteur et d'un effet observateur *(cf.* Swerts & Krahmer (2006) pour plus de précisions).

## *I.4. Importance du mouvement et de la dynamique, en particulier dans la perception des mouvements humains*

Une des caractéristiques fondamentales de notre environnement est qu'il est dominé par le mouvement. Ainsi, sa perception a une valeur adaptative, dans le sens où elle a valeur de survie chez les espèces : la perception rapide du mouvement (le danger d'un prédateur qui s'approche, ou, plus d'actualité, d'une voiture que vient vers nous) permet une réaction dans un délai bref.

Le mouvement ne fait pas qu'attirer notre attention. Il complète et organise également notre perception des objets et de leurs propriétés (en révélant par exemple la troisième dimension des objets), et nous renseigne sur les mouvements de notre propre corps lorsqu'il est mobile et qu'il doit interagir avec d'autres éléments de l'environnement (comme par exemple rattraper un ballon ou éviter un obstacle) (Streri, 2001, p. 164).

Le mouvement est non seulement une propriété de l'environnement que nous détectons, mais c'est également par lui que naît un « événement significatif ». En effet, la définition d'« événement » sous-entend à la fois les notions de changement et de mouvement.

D'autre part, les expériences de Johansson sur le mouvement biologique du corps humain montrent la façon dont le mouvement peut rendre plus claire une stimulation ambiguë. Il s'est intéressé à ce type très particulier de mouvement en utilisant la méthode des *point-lights,* c'est-à-dire des lumières attachées à certains points particuliers et précis du corps : aux articulations des membres et du cou. Il a ainsi montré que lorsque la personne portant les lumières ne bougeait pas, un observateur n'ayant à percevoir que les points lumineux, identifiait seulement un patron de lumières sans signification. Dès lors que la personne se mettait à marcher, le patron prenait alors sens et l'observateur identifiait en moins de 200ms un humain en train de marcher, même s'il ne s'agit perceptivement que de points lumineux, sans aucun autre détail de forme (Johansson, 1973).

Ces résultats insistent à la fois sur la spécificité du traitement perceptif du mouvement humain, dit biologique, et sur le lien existant entre perception du mouvement et perception de la forme. Cela implique deux flux différents de processus corticaux (*ibid*).

À la suite de Johansson, de nombreuses études ont été menées avec cette même méthode des points lumineux. Entre autres, les travaux de Cutting et collègues ont exploré la façon dont les observateurs sont capable de relier un patron particulier de points lumineux en mouvement à des propriétés telle que le genre (Barclay, Cutting, & Kozlowski, 1978) ou même l'identité (Cutting & Kozlowski, 1977) de la personne en

train de marcher portant les *point-lights*. Par la suite, Pollick (2004) a même identifié les paramètres des *point-lights* en mouvement, notamment temporels, qui sont utilisés pour identifier un « Human Movement style », voire même se servir de cette information pour reconnaître une personne .

Enfin, toujours en utilisant cette méthode, mais cette fois en l'adaptant pour étudier seulement les mouvements des visages, Bassili (1978 ; 1979), puis Bruce & Valentine (1988) ont testé la perception d'expressions faciales actées. Ils ont montré que les juges étaient capables de reconnaître très correctement ces expressions, simplement à partir du patron de déplacement de 50 ou 100 points lumineux positionnés au hasard sur le visage (concernant respectivement les études de Bassili et de Bruce et Valentine).

Plus particulièrement, concernant les mouvements liés aux expressions faciales émotionnelles, Kaiser, Wehrle, & Schmidt (1998) ont confronté deux positions théoriques au niveau de leur modélisation et de l'implémentation des paramètres temporels : celle d'Ekman et autres chercheurs partisans des émotions discrètes, et celle des évaluations cognitives de Scherer.
L'objectif était de vérifier si les expressions faciales étaient plutôt des patrons prototypiques de mouvements (d'AUs selon la terminologie d'Ekman), ou des mouvements singuliers reliés chacun (ou par groupe) à une évaluation cognitive particulière (pour plus de détails sur les deux positions théoriques, se reporter au Chapitre 1, et à Kaiser, Wehrle & Schmidt, 1998). Du point de vue de la dynamique, cela signifie que la dynamique du patron de mouvement est fixe dans le premier cas, alors qu'il dépend de la succession des évaluations cognitives dans le deuxième.
La comparaison des prédictions d'expressions faciales résultant de la modélisation de ces deux approches théoriques, a montré que les patrons prototypiques, décrits par Ekman et Friesen dans les prédictions d'émotions du manuel du FACS, ne sont produits que rarement. De plus, les résultats suggèrent qu'associer des AUs particulières aux différentes évaluations cognitives semble être une procédure appropriée pour générer les expressions faciales émotionnelles de manière dynamique (voir aussi Schmidt (1998), et plus récemment Scherer & Ellgring (2007)).

Tous ces travaux renforcent l'importance du mouvement dans la perception, mais laissent surtout apparaître le rôle fondamental de la dynamique des mouvements, en tout cas concernant les expressions faciales.

Dans la même idée, Cohn (2007) souligne l'importance des paramètres temporels de manière plus globale, à travers plusieurs exemples :

- deux ensembles d'actions faciales qui ont la même configuration, mais une organisation temporelle différente sont interprétées différemment ;

- la vélocité (*velocity*) et la durée de la première phase (*onset phase*) de sourires spontanés sont corrélées, contrairement aux sourires volontaires. Ainsi les sourires spontanés *vs.* volontaires peuvent être discriminés à près de 90%, rien qu'avec leurs paramètres dynamiques ;
- les expressions faciales subtiles sont identifiables seulement dans une présentation dynamique ;
- des expérimentations ont montré que des stimuli dynamiques apportaient plus d'informations que les mêmes stimuli présentés en multi-statique. Cela montre que ce n'est pas l'évolution du mouvement / comportement dans le temps qui apporte de l'information, mais bien la dynamique, c'est-à-dire les phases de vitesse et d'accélération.

Il reste à savoir comment se combinent les informations statiques et dynamiques. En effet, Humphreys, Donnelly, & Riddoch (1993) ont rapporté le cas d'un patient prosopagnosique (Bodamer, 1947) [68] incapable de reconnaître des expressions faciales à partir d'images statiques, mais qui arrive cependant à les identifier lorsqu'il a accès à des informations de mouvement. Ainsi, les expériences ne révèlent pas de détériorations de ses capacités lorsque les expressions sont produites sous ses yeux par l'expérimentateur, ni même lorsque les informations accessibles proviennent seulement du déplacement de points lumineux. Selon ces chercheurs, il s'agit d'un cas de dissociation entre deux processus : le premier impliqué dans la reconnaissance d'expressions faciales statiques ; le second dans la reconnaissance d'expressions faciales dynamiques. Ces résultats vont dans le même sens que ceux de (Desimone & Ungerleider, 1989), qui montrent de manière plus globale et à partir de données physiologiques, l'existence de systèmes fonctionnels séparés pour le traitement des formes en mouvement et celui des formes statiques. Généralisant leur résultats, Humphreys, Donnelly, & Riddoch (1993) suggèrent ainsi que les expressions émotionnelles sont perçues par des des canaux sensibles au mouvement et aux formes statiques, qui catégorisent chacun les expressions séparément. Les sorties de ces traitements seraient ensuite combinées à un niveau supérieur dans un système plus central par exemple pour une « évaluation sociale » (selon leurs propres termes).

[68] La prosopagnosie est un syndrome avec une atteinte de la capacité à reconnaître les visages familiers (Bodamer, 1947).

## *I.5. Enjeux théoriques des tests de perception / d'identification des états du Feeling of Thinking*

Les parties précédentes soulignent la multiplicité des traitements qui entrent en jeu concernant le visage : traitements du visage en tant que tel, avec d'un côté la reconnaissance du visage (information sur l'identité), de l'autre la perception des expressions faciales.

En parallèle, les expressions du visage et leur contexte (*e.g.* la gestualité les accompagnant ou même la scène dans laquelle la personne se trouve) est une forme comme une autre, avant d'être un élément particulier. C'est pourquoi il peut être supposé que les principes de la théorie de la *Gestalt* vont s'appliquer dans certains cas.

Une spatialisation de l'information propre au visage a pu être observée par certains auteurs. Dans un but technologique de synthèse, il n'est cependant pas souhaitable de générer uniquement les parties les plus informationnelles perceptivement. En effet, c'est potentiellement la méta-structure de l'expression qui est informationnelle. Ainsi, concernant l'additivité (spatiale mais aussi audio-visuelle dans cet exemple), si nous générons une expression combinant ce qui est le mieux reconnu en audio, et le mieux reconnu en visuel, cette expression ne sera pas forcément mieux reconnu que le naturel.

Enfin, étant donné que les expressions faciales font partie des mouvements biologiques, le visage va subir des traitements statiques comme dynamiques.

Ainsi, les informations concernant globalement le *Feeling of Thinking* peuvent être transmises selon différents paramètres visuels, et peut-être de manière spatialisée. Il s'agit ici de comprendre l'organisation de ce type d'information, à la fois au niveau modal et temporel, sans oublier le rôle possible de l'« effet observateur » dans sa perception.

# II. Méthodologie expérimentale

## *II.1. Objectifs de l'évaluation perceptive de nos icônes gestuelles*

Après la phase d'étiquetage exhaustive par les « experts », telle qu'elle a été décrite Chapitre 3 IV., nous avons cherché à évaluer perceptivement l'effet communicatif de certaines des Icônes Gestuelles relevées, auprès de juges naïfs. Par cette phase d'expérimentation, il s'agit d'établir des primitives communicatives. Ces primitives peuvent correspondre à des icônes ou des groupes d'icônes particuliers, comme à de simples paramètres, tel un type de dynamique dans le mouvement. Il s'agit également de tester les processus cognitifs qui entrent en jeu dans la perception de nos IGs, tels que ceux qui ont été décrits dans la première partie de ce chapitre.

Pour cela, des évaluations perceptives ont été menées : des stimuli ont été extraits à partir des signaux annotés, puis ont été donnés à percevoir à des juges naïfs. Les stimuli sont purement visuels pour cette évaluation perceptive, c'est-à-dire sans aucune production sonore, ni parole, ni autre son. La tâche demandée aux juges en perception était d'associer les stimuli à ce que les sujets disaient avoir ressenti à ce moment là, à leur auto-annotation.

Si les juges naïfs confirment perceptivement le choix d'auto-annotation des sujets, ou du moins s'ils évaluent les stimuli de manière cohérente entre eux, nous pouvons alors considérer les icônes testées comme pertinentes communicativement.

## *II.2. Choix des stimuli statiques et dynamiques : labels et IGs*

Nous avons d'abord sélectionné deux sujets (parmi les dix-sept), féminins, de culture française, et dont le comportement a été différent pendant l'expérimentation : F_T se dit particulièrement stressée par le scénario, et adhère au final à l'assertion que ses capacités cognitives sont détériorées, alors que F_S remet plutôt en question le logiciel.

Nous avons sélectionné pour chacun de nos sujets des labels d'auto-annotations de manière à ce qu'ils soient représentatifs de l'évolution de son état mental et affectif au cours du scénario. Ces labels (*cf.* Chapitre 3 III. pour les détails concernant cette partie de la méthodologie), servent de monnaie d'échange entre les sujets, s'exprimant lors des enregistrements, et juges observateurs lors de l'évaluation perceptive.

Nous ne prenons aucune décision quant à la l'interprétation objective selon une théorie, à la simplification ou à la généralisation de l'expression de l'auto-annotation, de manière également à ne pas décider *a priori* de ce qui est pertinent ou non dans

cette dernière. Cela signifie que les labels utilisés lors de l'évaluation, et donc les réponses / états de *FoT* proposés aux juges, sont les expressions utilisées par les sujets, telles qu'elles ou presque. Les seules modifications qui ont pu être apportées sont : le changement des pronoms de la première à la troisième personne ; le regroupement de deux labels d'auto-annotation de sens proche, noté « label 1 / label 2 » (*e.g.* « un peu perdue / perplexe »).

Ces différences entre auto-annotations données par les sujets et labels proposés aux juges sont récapitulés Annexe 4, accompagnées pour chaque auto-annotation de sa place approximative dans le scénario.

D'autres critères ont également guidé la sélection de ces auto-annotations :

- celles qui faisaient référence à un état simple non mélangé, et qui étaient exprimées dans une morpho-syntaxe simple. Leur intérêt était d'être parfois communes à plusieurs sujets, *e.g.* « stressée », « concentrée » ;
- des expressions simples comportant un modifieur, telles « **un peu** perdue » et « écoute **attentivement** » ;
- des expressions plus complexes, composées de plusieurs états, mais n'exprimant jamais d'expressions de type « mélange d'émotion ». Les expressions sélectionnées dans cette catégorie-là concernent toujours une combinaison de deux « sortes » d'états différents, par exemple un état mental et un état affectif (*e.g.* « envie de rigoler et répond au hasard ») ;
- une expression courante donnée par un des sujets : « ris jaune de mes résultats » ;
- une expression moins courante, mais dont le sujet a précisé ce qu'elle signifiait pour lui dans l'auto-annotation même : « "emprise" du logiciel, dans le sens où je suis les consignes du mieux que je peux ». Il s'agit là d'un état typique de l'interaction personne-machine. Il peut donc être utile de mesurer si l'explication cognitive naïve donnée par le sujet peut être discriminée par exemple de la concentration (dont il est peut-être une composante, ou du moins, dont il est du moins proche

Nous avons ainsi retenu dix labels d'auto-annotation pour le test concernant les icônes gestuelles du sujet T : « hésitante », « stressée », « mal à l'aise / inquiète », « angoissée / oppressée », « rassurée / plus détendue », « calme / va bien », « un peu perdue / perplexe », « déçue », « étonnée », « concentrée » ; et neuf pour celui du sujet S : « pas concentrée et envie de rigoler », « "ris jaune" de mes résultats », « écoute attentivement », « "emprise" du logiciel, dans le sens où je suis les consignes du mieux que je peux » (noté « "emprise" du logiciel » dans la suite de cet article), « stressée », « envie de rigoler et répond au hasard », « concentrée et répond au hasard », « concentrée » et « déçue ». Les labels d'auto-annotation utilisés étant évidemment

différents d'un sujet à l'autre, chaque sujet a donné lieu à une partie bien distincte lors des tests.

Parmi les données issues de l'étiquetage, nous avons sélectionné les stimuli de manière à ce qu'ils soient des icônes représentatives du comportement des sujets au cours du scénario et de leurs labels d'auto-annotation. Pour chaque label concernant le sujet T, quatre icônes gestuelles ont été sélectionnées, vs. deux icônes gestuelles concernant le sujet S.

Les stimuli du test peuvent être représentés schématiquement de cette manière :

1 icône gestuelle <-> 1 auto-annotation (« label »)

⇩

Plusieurs réalisations, Icônes Gestuelles, nous paraissant pertinentes, variant selon différents paramètres.

De plus, le choix des stimuli a été guidé par différents critères :

- les icônes où le sujet est en train de parler ont été éliminées, puisqu'en statique, il n'est pas possible de savoir si l'icône au niveau des lèvres est liée à ce qu'il dit ou si il s'agit d'expression de *FoT ;*
- globalement, les icônes sélectionnées sont soit relativement stables, soit ont au contraire une dynamique particulière (*e.g.* un mouvement ample, ou encore peu ample et furtif, etc.). Nous avons pour cela focalisé notre attention au niveau des lèvres, puisque c'est à cet endroit du visage que s'observe la parole (Chapitre 4 I.3.) ;
- les icônes non pertinentes méthodologiquement ont été supprimées : d'une part les mélanges d'autres icônes (primitives) et les transitions entre expressions de *FoT* ; d'autre part celles dont le lien avec l'auto-annotation du sujet n'est pas intuitif, et d'autant moins si cette auto-annotation fait référence à un état complexe ;
- la sélection finale des stimuli a enfin été réalisée intuitivement, de manière à obtenir le même nombre de stimuli pour un même label et entre les labels, afin de faciliter les traitements statistiques futurs.

Nous avons ainsi fait en sorte de tester des icônes gestuelles variées, en privilégiant celles qui nous semblait particulièrement pertinente vis-à-vis de l'expression des états du *FoT*. La liste des stimuli du test se trouve Annexe 5.

De plus, nous cherchons à comparer l'efficacité des icônes dynamiques *vs*. statiques pour transmettre l'information indiquée par les labels d'auto-annotation. pour cela, nous avons extrait une image statique supposée caractéristique, à l'apex du mouvement lorsque cela était possible, pour chaque icône gestuelle minimale sélectionnée comme stimulus (vidéos longues de 0,5 à 4 secondes).

## *II.3. Conditions de présentation haut / bas / entier*

Dans notre domaine les principales études concernent les émotions. De plus, il a été montré par différents travaux autour de la théorie d'Ekman (*cf.* Chapitre 4 I.3.) que le visage était très informatif, et que le haut et le bas du visage ne portaient pas les mêmes informations. C'est pourquoi nos stimuli ont été testés dans trois conditions de présentation : haut du visage seul (« haut »), bas du visage seul (« bas » Figure 19) et visage entier (« entier »).

Cela était d'autant plus important que nous nous replacerons plus tard chez le sujet parlant, et que les études en *eye-tracking* montrent une prédominance alternée de deux grandes zones : celle des yeux et celle de la bouche.

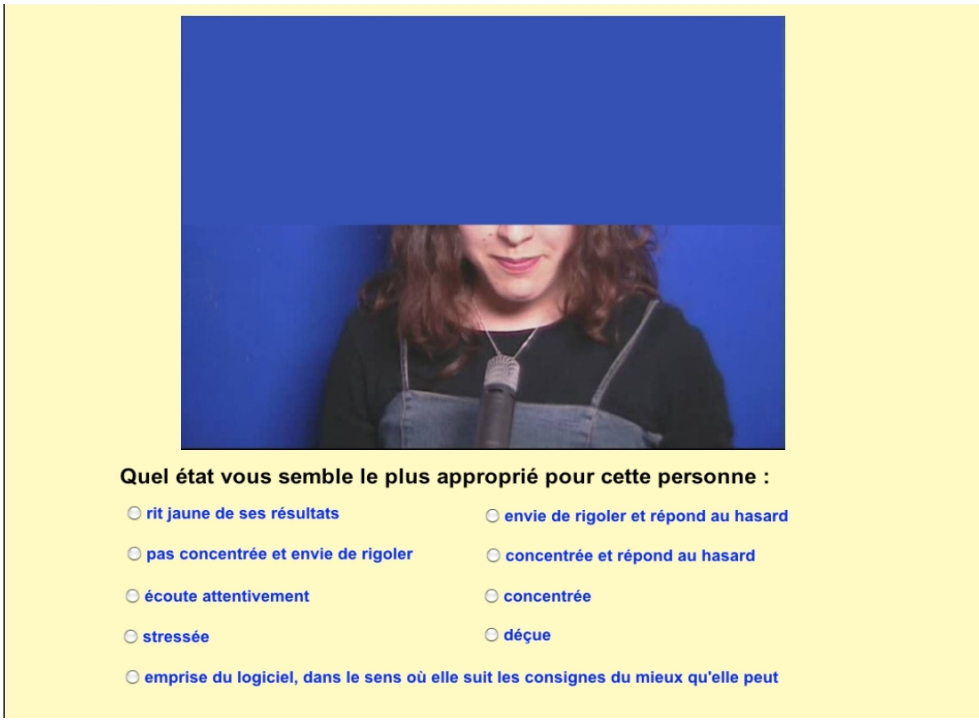



Figure 19: Exemple d'interface, avec un stimulus du sujet S en condition « bas »

Pour générer les stimuli des conditions « haut » et « bas », nous avons superposé manuellement, avec l'application Microsoft Powerpoint pour les stimuli statiques, un rectangle de la couleur du fond (bleu) pour chacun des stimuli. Les limites de ces rectangles sont les bords latéraux, et supérieur ou inférieur de l'image, et une ligne parallèle passant toujours par le haut des ailes du nez, bas du muscle zygomatique (Figure 19). Cette dernière a été définie de façon à cacher les mouvements de la mâchoire lors de la condition « haut » (bas du visage caché et haut visible). Dans le cas de stimuli dynamique, cette limite du rectangle évoluait avec la vidéo.

## *II.4.Déroulement des tests et ordre des stimuli*

Deux tests similaires ont ainsi été passés par seize juges naïfs : le premier avec les stimuli statiques, le second avec les icônes gestuelles proprement dites (les stimuli dynamiques). Chaque stimulus a été présenté en ordre aléatoire, une fois dans chaque condition de présentation, cela pour chacune des deux sessions du test perceptif, soit 120 stimuli pour le sujet T (10 labels * 4 stimuli par label * 3 conditions) et 54 pour le sujet S (9 labels * 2 stimuli par label * 3 conditions).

L'ordre de présentation des stimuli est important dans la mesure où un biais peut être rapidement introduit en inter-sujets, du fait de l'habituation aux visages des stimuli. Pour éviter cela, la moitié des sujets a commencé par le test de S., puis a continué par celui de T. ; l'autre moitié a fait l'inverse.

De plus, étant données les trois conditions de présentation des stimuli, la moitié des sujets s'est vu présenter d'abord le bas du visage, puis le haut, et enfin le visage en entier ; l'autre moitié a commencé par le haut du visage, puis le bas, et toujours le visage entier en dernier.

Les tests perceptifs ont consisté pour les juges en une tâche d'association entre les stimuli et un choix fermé (cases à cocher) parmi les labels d'auto-annotation du sujet (Figure 19).

Le test a été conçu avec l'application « Runtime Revolution Studio », en se fondant sur les scripts et l'interface d'un test dont le protocole était presque identique, réalisé par Albert Rilliard. Les scripts modifiés, qui définissent l'interface et permettent de présenter les stimuli, tout en enregistrant automatiquement les réponses des juges sous forme de fichiers texte se trouvent en Annexe 6.

Alors que le temps d'exposition n'était pas limité pour les stimuli statiques, les stimuli dynamiques pouvaient être rejoués pendant huit secondes. Ainsi, pour le test des stimuli statiques, l'interface n'était composée que d'un seul écran, contenant à la fois le stimulus, les cases à cocher pour les réponses et un bouton « suivant » pour passer au stimulus suivant. La même interface a été utilisée pour le test des stimuli dynamiques, avec un bouton « rejouer » remplaçant le bouton « suivant » tant que le juge pouvait rejouer la vidéo. Ce délai de 8 secondes a été choisi de manière à ce que les stimuli les plus longs (un peu plus de 4 secondes) puissent être rejoués une fois, pendant que les stimuli-vidéos les plus courts (environ une demi seconde) puissent être rejoués plusieurs fois, mais avec un temps de présentation maximal aux juges identiques quel que soit la durée du stimulus.

Malgré ces contraintes temporelles données par la conception même du test, il était demandé au juge de répondre aux stimuli le plus rapidement et le plus spontanément possible.

## *II.5. Sélection des juges pour le test perceptif et explications à leur fournir au préalable*

Dix-sept et quinze juges (respectivement pour les tests de T et de S) ont vu leurs réponses exploitées dans les résultats. Tous étaient de langue maternelle française. Ils ont aussi été choisis d'âge varié et avec une répartition par sexe assez équilibrée.

Un autre facteur à ne pas négliger quant au choix des juges était leur familiarité ou non avec chacun des deux sujets : il est en effet probable que ce facteur ait une influence sur l'identification des expressions de *FoT* d'autrui, et par suite sur les taux de bonne association entre stimuli et labels d'auto-annotation. Nous avons donc privilégié des juges ne connaissant aucun des deux sujets, et noté dans le cas contraire quel(s) sujet(s) étaient connus des juges. Cette information, ainsi que le sexe et l'âge des juges ont été enregistrées automatiquement.

Par ailleurs, il est important de toujours garder à l'esprit que nos travaux se trouvent dans de la « communication située », puisqu'une interaction ne peut avoir lieu hors contexte. Toute personne, participant ou observateur d'une situation de communication, a en effet une perception cognitive du contexte. Par conséquent, il était nécessaire de préciser aux juges passant les tests de perception la situation dans laquelle se trouvent les sujets (interface du test où est présentée la situation Annexe 7), c'est-à-dire devant un écran sur lequel ils reçoivent des consignes à certains moments, qu'ils doivent exécuter, après quoi ils obtiennent leurs résultats.

# III. Résultats

## *III.1. Prétraitements et graphes de confusions*

Afin de traiter les résultats individuels des juges, nous avons adapté à nos données un script Java qui avait été développé par Nicolas Audibert pour traiter des données formatées de façon similaire. Ce programme regroupe les résultats sous la forme d'une matrice de confusion par juge et par condition (haut / bas / entier). Il enregistre également les résultats, convertis en bonne ou mauvaise réponse avec un format compatible avec le logiciel SPSS.

Nous avons ensuite calculé les matrices de confusion globales en additionnant les résultats individuels des juges dans chacune des conditions, à l'aide du logiciel Microsoft Excel. Nous avons enfin calculé en pourcentage les taux de réponses selon le label présenté, et ainsi obtenu les matrices de confusion, sur lesquelles seront fondées les analyses qui suivront (voir un exemple Tableau 3). Ces matrices de confusions ont été réalisées pour chacun des sujets et pour chacun des tests (avec stimuli statiques ou dynamiques, et conditions de présentation « haut » / « bas » / « entier »). Elles permettent de conserver toutes les informations concernant la distribution des réponses attribuées aux stimuli de chaque catégorie (*i.e*. le s labels).

À partir de ces matrices, nous avons généré des graphes de confusions. Nous avons, pour la représentation graphique, fixé un seuil, à deux fois le seuil du hasard (c'est-à-dire 20% pour le sujet T -10 labels, donc le seuil du hasard est à 10%-, et 22% pour le sujet S -9 labels-). Seuls les reports ou confusions entre labels supérieurs à ce seuil ont été représentés par des flèches reliant les labels en question. Nous appelons « report » le cas où les réponses données par les juges à des stimuli correspondant à un label ont été reportés sur un autre label (*e.g.* les cases en gris moyen dans le tableau 3). Quant aux confusions, ils s'agit pour nous d'un report mutuel de réponses concernant deux labels.

C'est également ce seuil qui nous a fait classé les labels en « bien reconnus » (*e.g.* les cases en gris foncé de la diagonale dans le tableau 3), ou « mal reconnus » (*i.e.* reconnu à un taux inférieur à ce seuil, en gris clair sur la diagonale de la matrice -Tableau 3-). Voir les différentes représentations des labels selon leur taux de reconnaissance Figure 20, et le graphe de confusion issu de la matrice de confusion du Tableau 3 avec un exemple d'interprétation d'éléments représentés Figure 21.

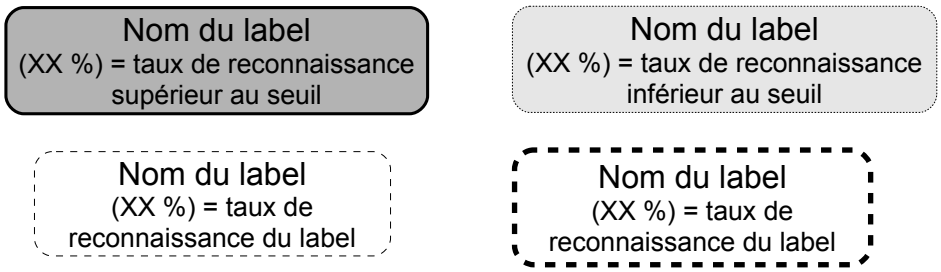



Figure 20: Représentations des labels dans les graphes de confusion: en haut à gauche, reconnu au dessus de notre seuil; en haut à droite, reconnu au dessous de notre seuil; en bas à gauche, label dont la répartition des réponses données à ce label n'est pas différente du hasard (selon Khi2, avec p>0,01); en bas à droite, idem en bas à gauche, mais avec p>0,05.

Ensuite, nous avons cherché à établir des méta-classes, c'est-à-dire des groupements de labels dont le taux de reconnaissance global est élevé (plus de 40% de reconnaissances correctes en condition « entier »). Toutefois, au sein de ces groupements, chacun des labels, pris indépendamment, peuvent avoir un taux de reconnaissance faible, et / ou un grand nombre de report de réponses, principalement sur les autres labels faisant partie du groupement. Nos méta-classes ont généralement émergé d'elle-même lors de la construction des graphes de confusions en condition « entier » : elles apparaissent comme des groupes de labels isolés, dont les reports de réponse se font presque exclusivement entre eux (*cf.* graphe de confusion de la Figure 21). Nous les avons notées par des ellipses dans les graphes de confusions. Nous

Tableau 3: Exemple de matrice de confusion
(Sujet T, stimuli statiques, condition "entier").
Les labels des stimuli se trouvent verticalement, les labels des réponses horizontalement

| F_T statique Entier | un peu perdue / perplexe | déçue | calme / va bien | mal à l'aise / inquiète | hésitante | stressée | rassurée / plus détendue | étonnée | angoissée / oppressée |
|---|---|---|---|---|---|---|---|---|---|
| un peu perdue / perplexe | 22,1% | 14,7% | 1,5% | 19,1% | 10,3% | 7,4% | 1,5% | 10,3% | 4,4% |
| déçue | 14,7% | 33,8% | 7,4% | 2,9% | 1,5% | 1,5% | 7,4% | 23,5% | 5,9% |
| calme / va bien | 7,4% | 1,5% | 17,6% | 0,0% | 2,9% | 0,0% | 51,5% | 1,5% | 0,0% |
| mal à l'aise / inquiète | 16,2% | 4,4% | 0,0% | 19,1% | 25,0% | 11,8% | 0,0% | 5,9% | 1,5% |
| hésitante | 19,1% | 2,9% | 5,9% | 10,3% | 17,6% | 2,9% | 16,2% | 16,2% | 1,5% |
| stressée | 16,2% | 11,8% | 0,0% | 10,3% | 20,6% | 11,8% | 0,0% | 5,9% | 4,4% |
| rassurée / plus détendue | 0,0% | 0,0% | 50,0% | 2,9% | 0,0% | 0,0% | 32,4% | 10,3% | 0,0% |
| étonnée | 26,5% | 13,2% | 0,0% | 5,9% | 7,4% | 8,8% | 2,9% | 11,8% | 4,4% |
| angoissée / oppressée | 16,2% | 1,5% | 4,4% | 25,0% | 11,8% | 19,1% | 2,9% | 7,4% | 11,8% |

avons ensuite calculé leur taux de reconnaissance global et les reports globaux entre elles, c'est-à-dire en considérant la méta-classe comme un seul et même label. Les reports ont seulement été notés sur les graphes lorsqu'ils étaient supérieur au seuil du hasard, soit 25% pour le sujet T (puisque nous avons relevé 4 méta-classes). Les taux de reconnaissance et reports sont représentés par les pourcentages et flèches en gras sur les graphes de confusions (Figure 21).

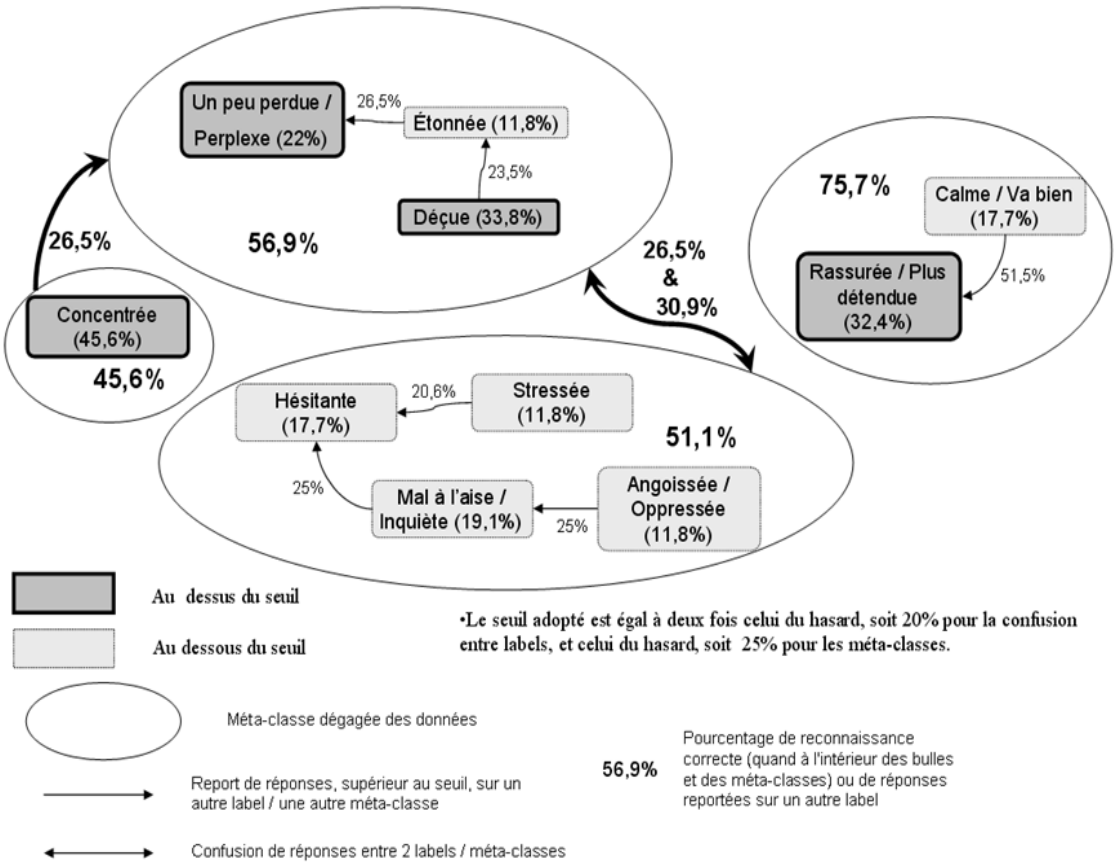



Figure 21: Exemple de graphe de confusion issu de la matrice de confusion -Tableau 3- (Sujet T, stimuli statiques, condition « entier ») et exemple d'interprétation : les stimuli correspondant au label « déçue » ont été correctement identifiés dans 33,8% des cas, soit au dessus de notre seuil de deux fois celui du hasard. Toutefois, 23,5% des réponses à ces mêmes stimuli se sont reportées sur « Etonnée », *i.e.* ces stimuli (auto-annotés « déçue » par le sujet lui-même) ont été identifiés comme « Etonnée » par les juges.

Pour les conditions « haut » et « bas », le graphe généré pour la condition entier correspondant (au niveau du sujet, et du type de stimuli -statiques ou dynamiques-) a servi de base de comparaison pour les résultats, en particulier pour les méta-classes (*i.e.* nous n'avons pas cherché de nouvelles méta-classes, mais avons simplement vérifié si la ou les méta-classes relevée(s) en condition « entier » était toujours pertinente.

Une grande partie des analyses effectuées a consisté en des analyses qualitatives et descriptives, fondées sur ces représentations graphiques des matrices de confusions.

## *III.2. Analyse statistique de chacun des tests perceptifs*

Une ANOVA à mesures répétées a été effectuée avec le logiciel de statistiques SPSS sur les données transformées pour chaque partie du test. Les facteurs testés sont : au niveau intra-juge, la condition de présentation, avec trois niveaux (haut et bas du visage seuls, et visage entier), et les labels de réponses ; au niveau inter-juges, le sexe et l'âge des juges, ainsi que leur connaissance des sujets. Il apparaît que ni le sexe, ni l'âge, ni le fait de connaître ou non le sujet, n'ont de répercussions sur l'identification des stimuli. Le seul effet significatif ($p<.01$, pour chacun des deux sujets, et à la fois avec les stimuli statiques et dynamiques) est celui des labels de réponse associés aux stimuli présentés.

Des tests de Khi-deux ont également été réalisés sur les matrices de confusion (par rapport à des matrices contenant les taux correspondant au seuil du hasard). Un test de Khi-deux compare la distribution des réponses observée à une distribution théorique des réponses, dans le cas où les juges répondraient au hasard. Ce test nous a donc permis de prendre en compte les confusions dans la significativité des résultats, et non, comme dans le test d'ANOVA, seulement les réponses correctes. Nous avons ainsi relevé des cas de labels dont la distribution des réponses par les juges n'était pas significativement différente du hasard (soit à $p>0,05$, soit à $p>0,01$).

Ces labels sont représentés tel qu'indiqué dans la Figure 20 dans nos graphes de confusions.

Enfin, en utilisant le logiciel R[69], nous avons généré des *clusters* hiérarchiques, toujours pour chacun des sujets et des tests (avec stimuli statiques / dynamiques) et chacune des conditions de présentation.

Leur interprétation a fait émerger plusieurs *clusters*, différents selon la condition de présentation des stimuli, et selon leur nature statique ou dynamique. Les résultats de l'analyse de quelques *clustering* sont présentés dans la partie III.5.

[69] Pour plus d'information sur le projet R, voir http://www.r-project.org/

## *III.3. Test des stimuli statiques*

### III.3.1. Résultats spécifiques au sujet T

Globalement, dans toutes les conditions, c'est le label « concentrée » qui est le mieux reconnu (de 39,7% à 52,9% de reconnaissance correcte). Nous pouvons également relever qu'il attire les réponses des autres stimuli, notamment en condition « haut » et « bas ».

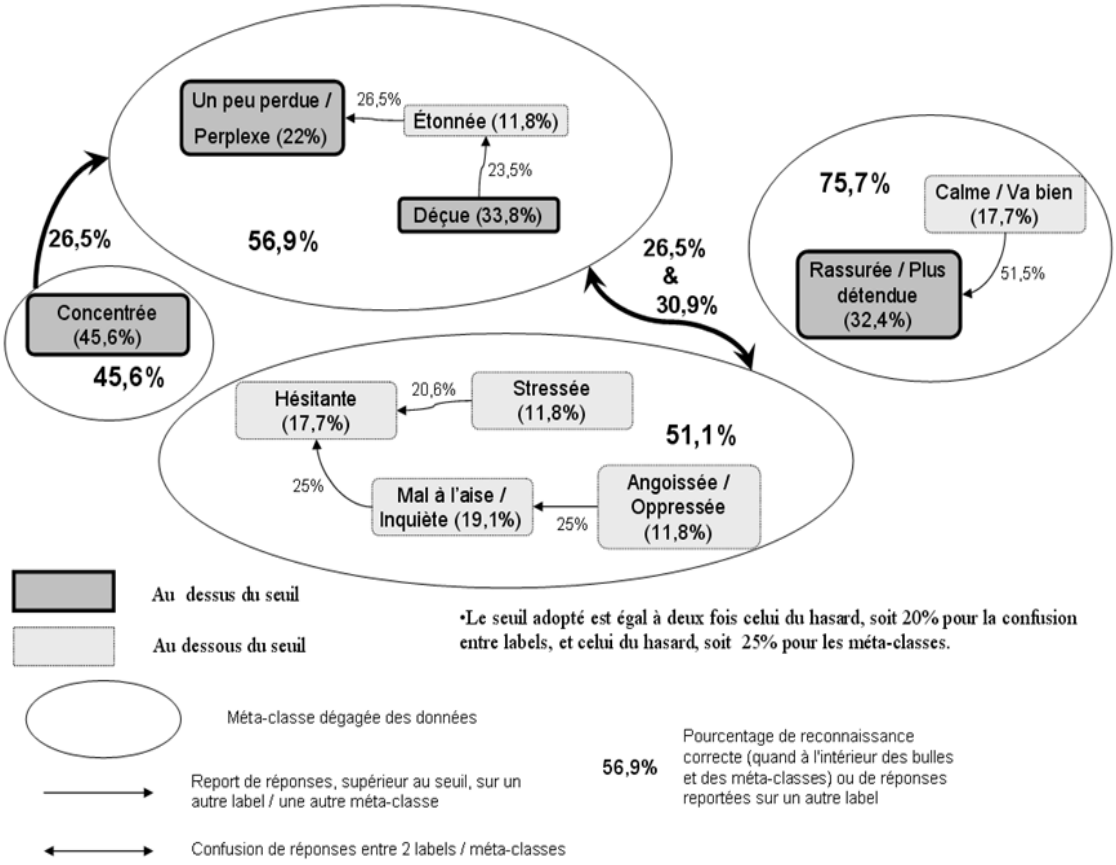



Figure 22: Résultats pour les stimuli statiques en condition « entier » - Sujet T
les labels « concentrée », « un peu perdue / perplexe », « déçue » et « rassurée / plus détendue » ont été reconnus au dessus du seuil de deux fois le niveau du hasard. Les autres labels ont eu des reports de réponses sur d'autres labels.

Les reports et confusions entre labels de la condition « entier » (*cf.* Figure 22) nous ont permis de dégager quatre méta-classes pour la matrice de confusions de la condition entier :

- « rassurée / plus détendue » et « calme / va bien » (75,7 % de taux de reconnaissance correct pour la méta-classe) ;
- « hésitante », « stressée », « mal à l'aise / inquiète » et « angoissée / oppressée » (51,1%) ;

- « un peu perdue / perplexe », « déçue » et « étonnée » (56,9%) ;
- « concentrée » seul (45,6%).

En conditions « haut » (Figure 23) et « bas » (Figure 24), les quatre méta-classes sont restées bien reconnues, mais avec des confusions mutuelles plus importantes qu'en condition « entier ».

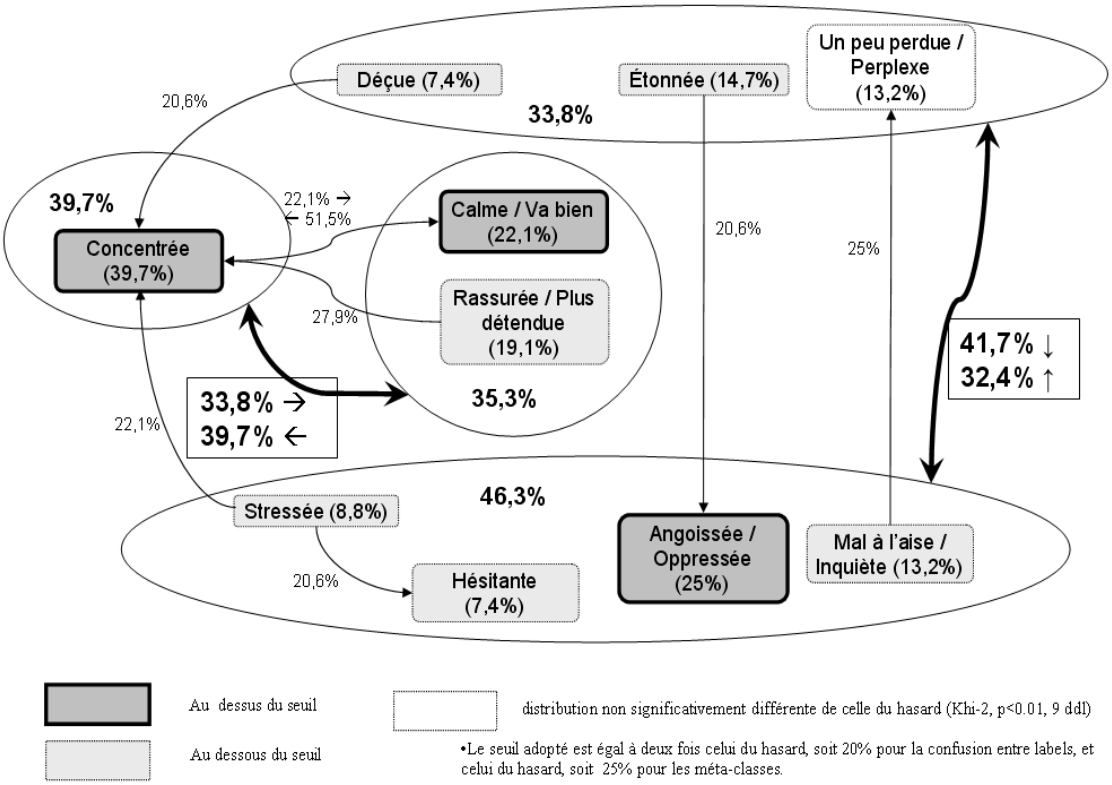



Figure 23: Résultats pour les stimuli statiques en condition « haut » - Sujet T
complément de légende identique à celle de la figure précédente.

Les méta-classes sont ici confondues deux à deux :
- « étonnée + déçue + un peu perdue/perplexe » avec « mal à l'aise/inquiète + hésitante + stressée + angoissée/oppressée » ;
- « calme/va bien + rassurée/plus détendue » avec « concentrée »).

En condition « bas » (Figure 24), nous retrouvons la même confusion, mais moindre qu'en condition « haut », entre « étonnée + déçue + un peu perdue » et « mal à l'aise/inquiète + hésitante + stressée + angoissée/oppressée ». « concentrée » est quant à elle toujours bien reconnue, et même mieux qu'en condition entier si nous nous fions seulement aux chiffres. Les informations concernant la concentration semble donc se trouve donc majoritairement dans le bas du visage pour ce sujet.

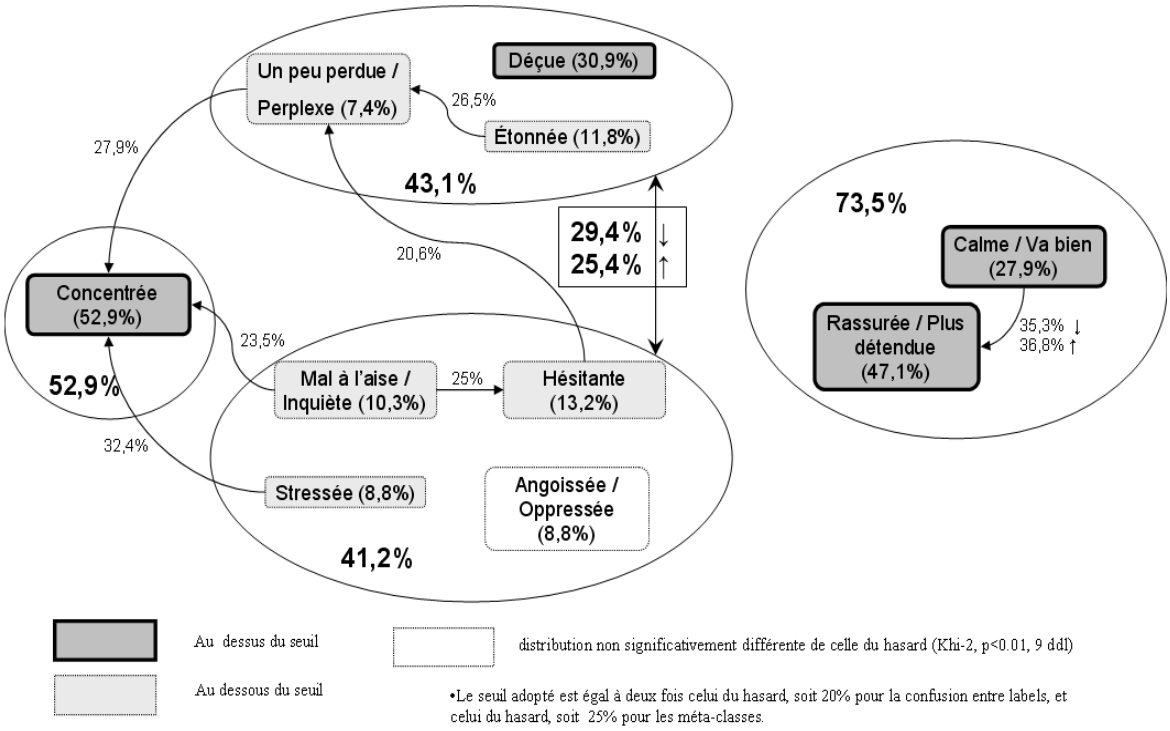



Figure 24: Résultats pour les stimuli statiques en condition « bas » - Sujet T
(complément de légende identique à celle de la figure 22)

La distribution des réponses pour les labels « un peu perdue/perplexe » en condition « haut » et « angoissée/oppressée » en condition « bas » n'est pas significativement différente de celle du hasard (Khi-2, p>0.01, 9 ddl), « angoissée/oppressée » étant curieusement le mieux reconnu en condition « haut », et non « entier ». « concentrée » est quant à lui le mieux reconnu (de 39,7% « haut », à 52,9% « bas ») et a tendance à attirer les autres réponses.

### III.3.2. Résultats spécifiques au sujet S

En ce qui concerne le sujet S, les labels « pas concentrée et envie de rigoler » pour les conditions « entier » et « bas », et « concentrée et répond au hasard », « "emprise" du logiciel » et « concentrée » pour la condition « haut » n'obtiennent pas une distribution significativement différente de celle du hasard (Khi-2, p>0.01, 8 ddl). Comme pour le sujet T, le label « concentrée », bien reconnu en condition « entier », a tendance à attirer vers lui les réponses données aux stimuli autres que ceux de la méta-classe (*cf.* Figure 25).

En condition « entier », et comme le label « concentrée », « écoute attentivement » et « envie de rigoler et répond au hasard » ont été reconnus au dessus du seuil de deux fois le niveau du hasard.

Par ailleurs, le méta-classement a été moins productif que pour le sujet T, puisque nous avons seulement dégagé une méta-classe : « envie de rigoler et répond au hasard », « pas concentrée et envie de rigoler » et « "rit jaune" de ses résultats » (taux d'identification alors de 61,5% en condition « haut » à 72,9% en condition « entier »).

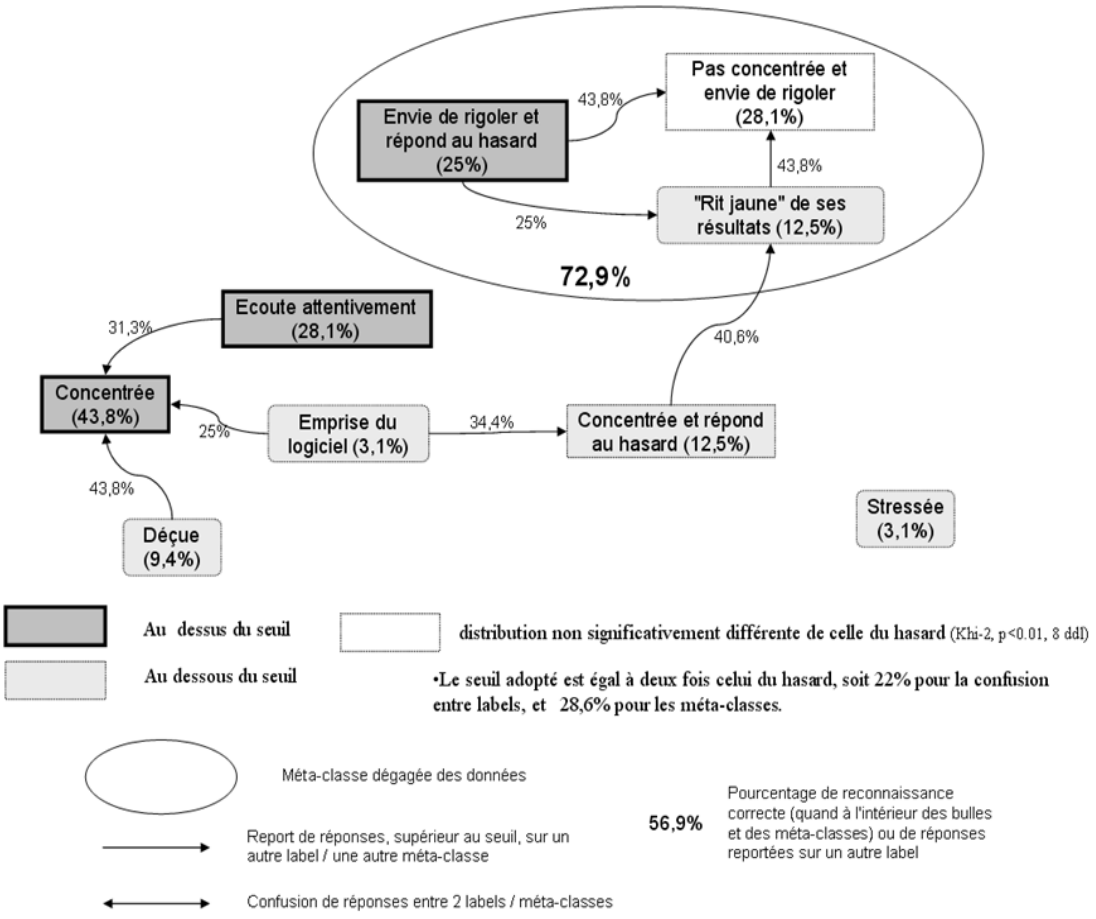



Figure 25: Résultats pour les stimuli statiques en condition « entier » - Sujet S.

Nous pouvons également observer que « concentrée et répond au hasard » se reporte sur la méta-classe définie, et en particulier sur « "rit jaune" de ses résultats »

En condition « bas » (Figure 26) et comme pour la condition « entier », « concentrée » attire de nombreuses réponses et la méta-classe dégagée reste bien reconnue (à près de 70%). Toutefois, seul « envie de rigoler et répond au hasard » reste reconnu au dessus de notre seuil. De plus le label « "rit jaune" de ses résultats » est très mal reconnu et n'est la cible d'aucun report de réponses important. Il est également intéressant de constater que : « déçue » est non plus reportée sur « concentrée », mais sur « écoute attentivement » (qui lui même a ses réponses qui se reportent sur « concentrée » ; que « concentrée et répond au hasard » voit toujours ses réponses se reporter sur la méta-classe, mais cette fois sur « envie de rigoler et répond au hasard » en particulier (label avec qui il partage la notion de « réponse au hasard ».

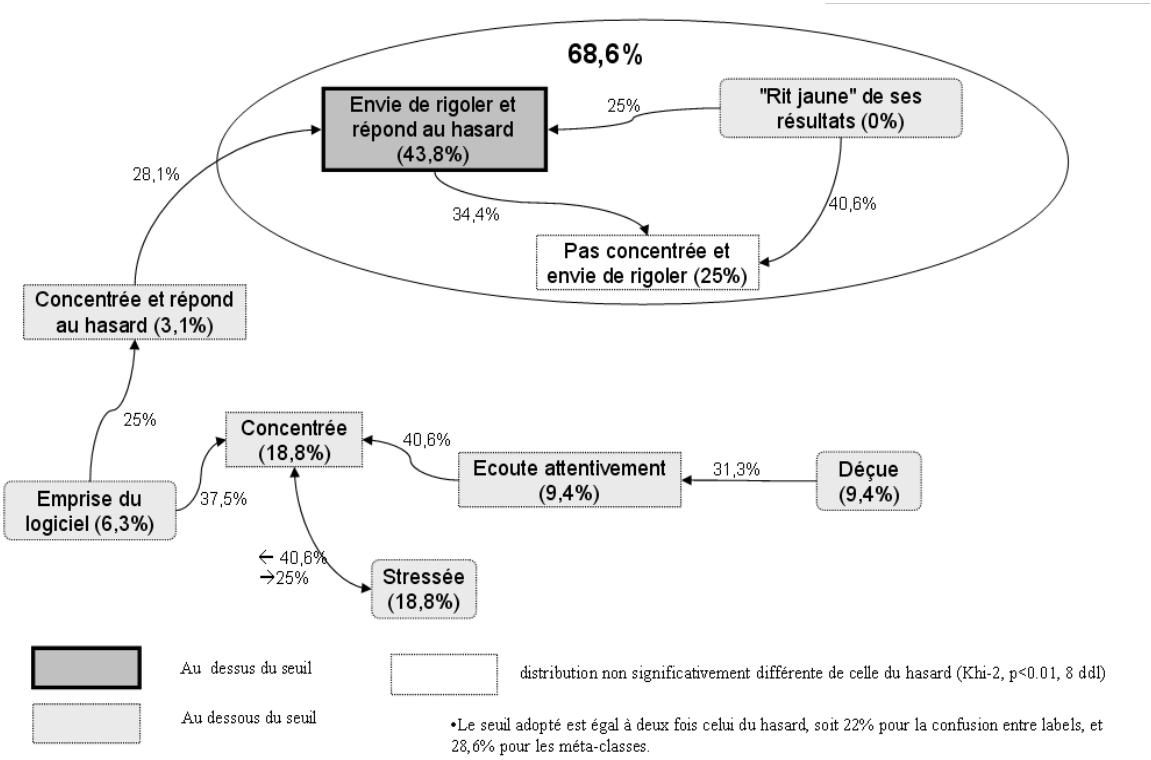



Figure 26: Résultats pour les stimuli statiques en condition « bas » - Sujet S
(complément de légende identique à celui de la figure 25)

De la même manière qu'en condition « bas », nous observons toujours mêmes grandes tendances en condition « haut » (Figure 27) qu'en condition « entier ». La différence réside cette fois-ci dans les réponses dont la distribution est proche du hasard concernant les stimuli « concentrée et répond au hasard », « "emprise" du logiciel » et « concentrée ». Ainsi, alors que ce dernier label attire les réponses aux stimuli des autres labels et est reconnu au dessus du seuil fixé (avec 31,3%), les réponses à ses propres stimuli sont données d'une manière proche du hasard.

D'autre part, l'identification des stimuli du label « écoute attentivement » passe curieusement de 28,1% à 43,8% de la condition « entier » à la condition « haut » (alors qu'il est reconnu en dessous du seuil en condition « bas »). Cela suggère que les informations concernant ce label se trouvent principalement dans le haut du visage pour ce sujet. Autre curiosité, « envie de rigoler et répond au hasard » est mieux reconnu dans les conditions « haut » et « bas » (avec respectivement 50% et 43,8% de bonne identification) qu'en condition « entier » (25%), pour laquelle ses réponses se reportent sur les autre labels de la méta-classe.

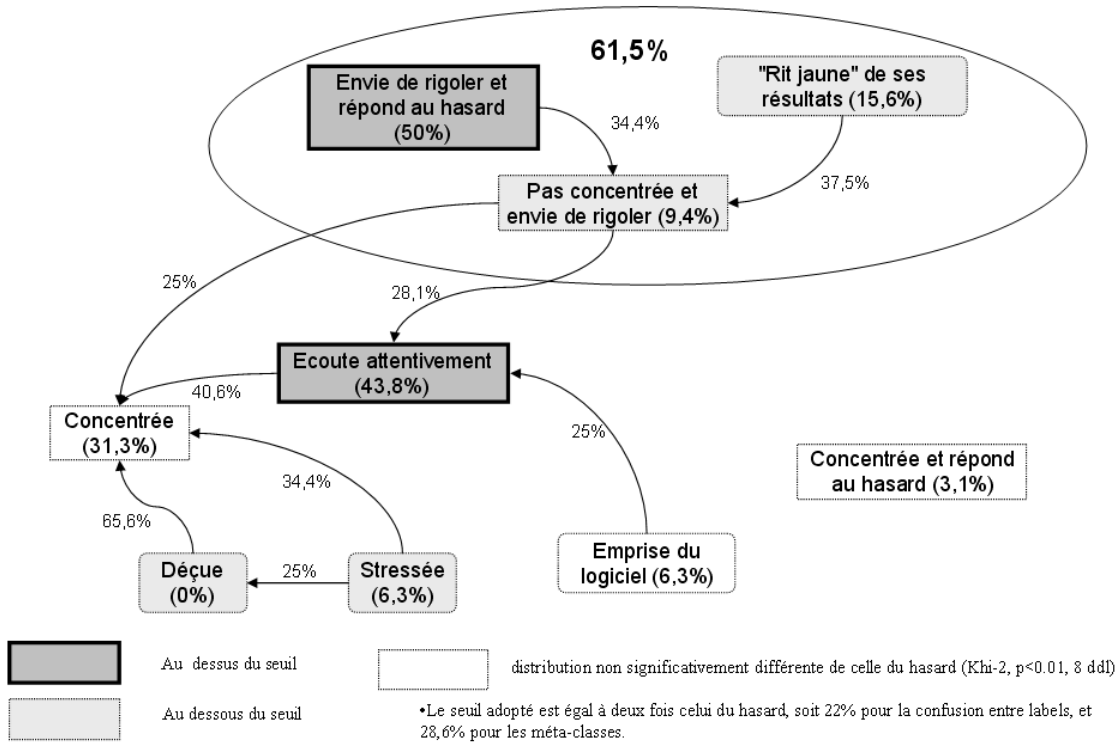



Figure 27: Résultats pour les stimuli statiques en condition « haut » - Sujet S
(complément de légende identique à celui de la figure 25)

### III.3.3. Inter-sujets

Concernant la comparaison des résultats en inter-sujets, peu de labels sont directement communs aux sujets. Toutefois il est à noter que chez nos deux sujets le label « concentrée » est le mieux reconnu dans toutes les conditions pour T, mais seulement en condition « entier » pour S. Quoi qu'il en soit, il a tendance à attirer les réponses des autres labels dans tous les cas (pour les deux sujets et dans toutes les conditions. D'autre part « déçue » est bien reconnu uniquement pour T, et seulement en condition « entier » et « bas ». Il est également intéressant de noter que pour les deux sujets, « concentrée » attire les réponses de « déçue » en condition « haut » à hauteur de 27,9% des réponses pour T et 65,6% pour S. « stressée » n'est quant à lui jamais reconnu au dessus de notre seuil (en tout cas sur les icônes sélectionnées et présentées), quels que soient le sujet et la condition. Cela signifie soit que les icônes présentées n'ont pas été représentatives du label, soit que l'information concernant ce label n'est pas transmise par une photo, *i.e.* par une modalité visuelle statique.

## *III.4. Test des stimuli dynamiques*

### III.4.1. Résultats spécifiques au sujet T.

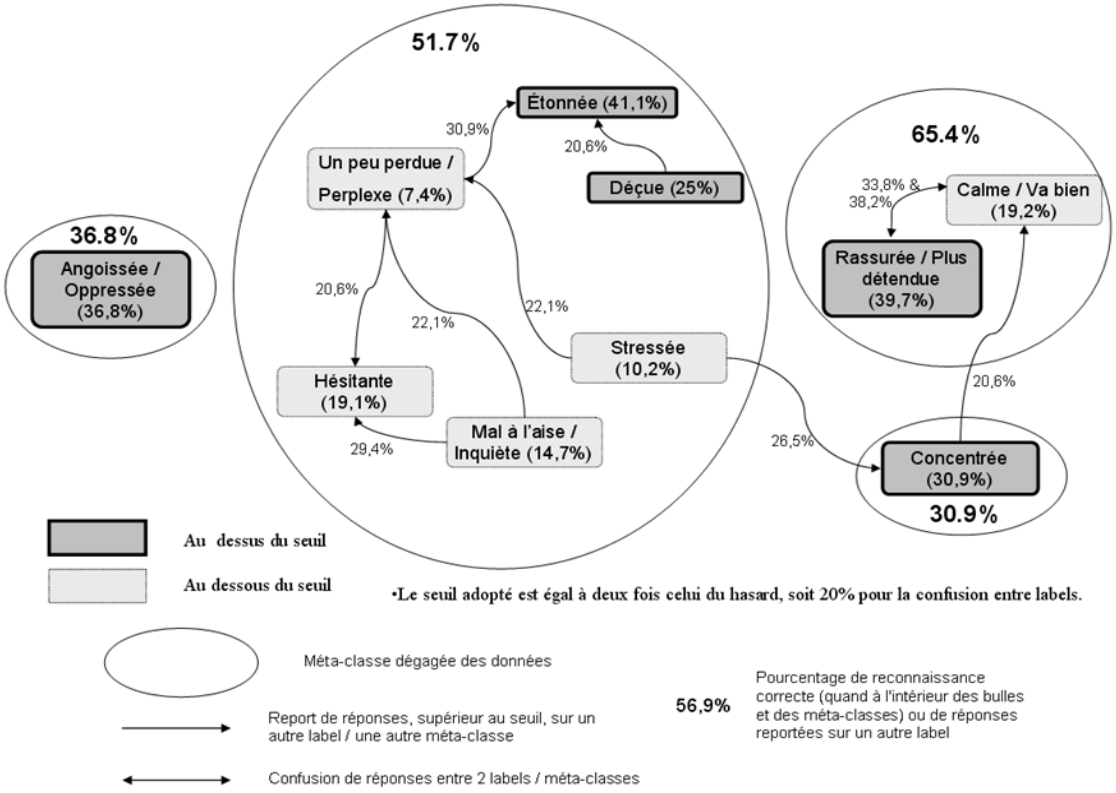



Figure 28: Résultats pour les stimuli dynamiques en condition « entier » - Sujet T

Les labels « concentrée », « étonnée », « déçue », « angoissée / oppressée » et « rassurée / plus détendue » ont été reconnus au dessus du seuil de deux fois le niveau du hasard.

En condition « entier », les reports et confusions entre labels nous ont permis de dégager quatre méta-classes (*cf.* Figure 28) :

- « rassurée / plus détendue » + « calme / va bien » (65,4 %) ;
- « hésitante » + « stressée » + « mal à l'aise / inquiète » + « un peu perdue/ perplexe » + « déçue » + « étonnée » (51,7%), avec « un peu perdue / perplexe » mal identifié, mais qui attire les réponses de tous les autres labels de la méta-classe, sauf de « déçue » (dont les réponses se reportent sur « étonnée » ;
- deux méta-classes de labels isolés : « angoissée / oppressée » (36,8%) et « concentrée » (30,9%).

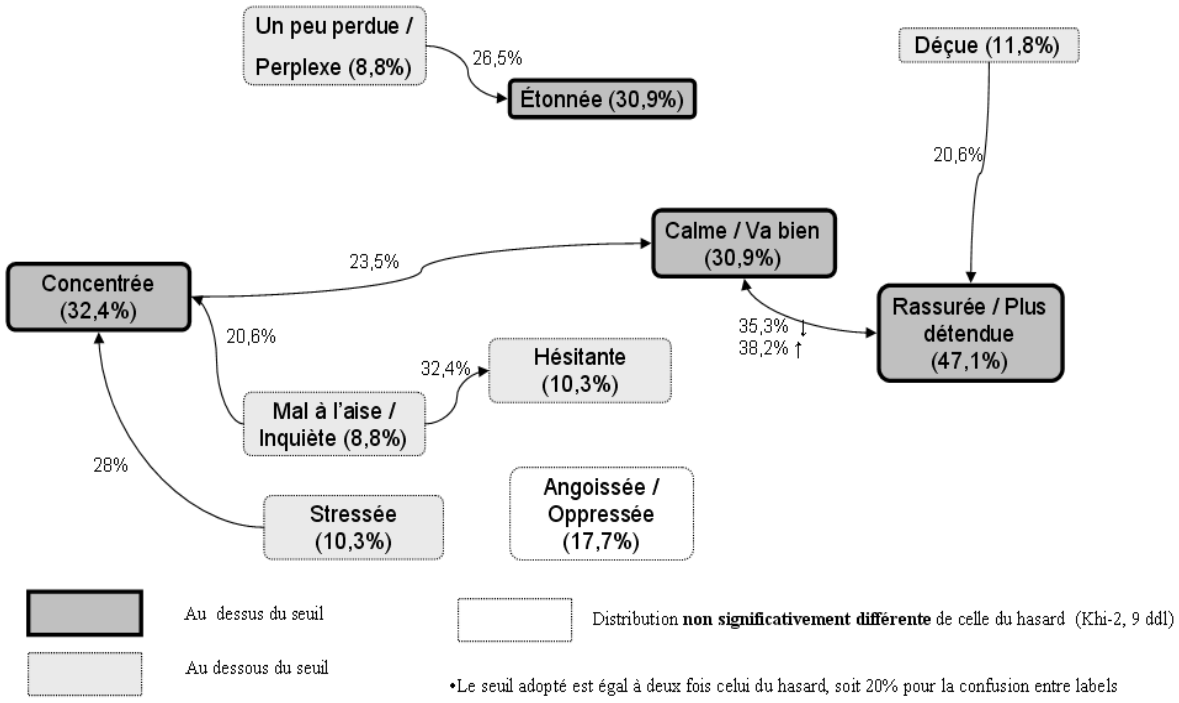



Figure 29: Résultats pour les stimuli dynamiques en condition « bas » - Sujet T
(complément de légende identique à celui de la figure 28)

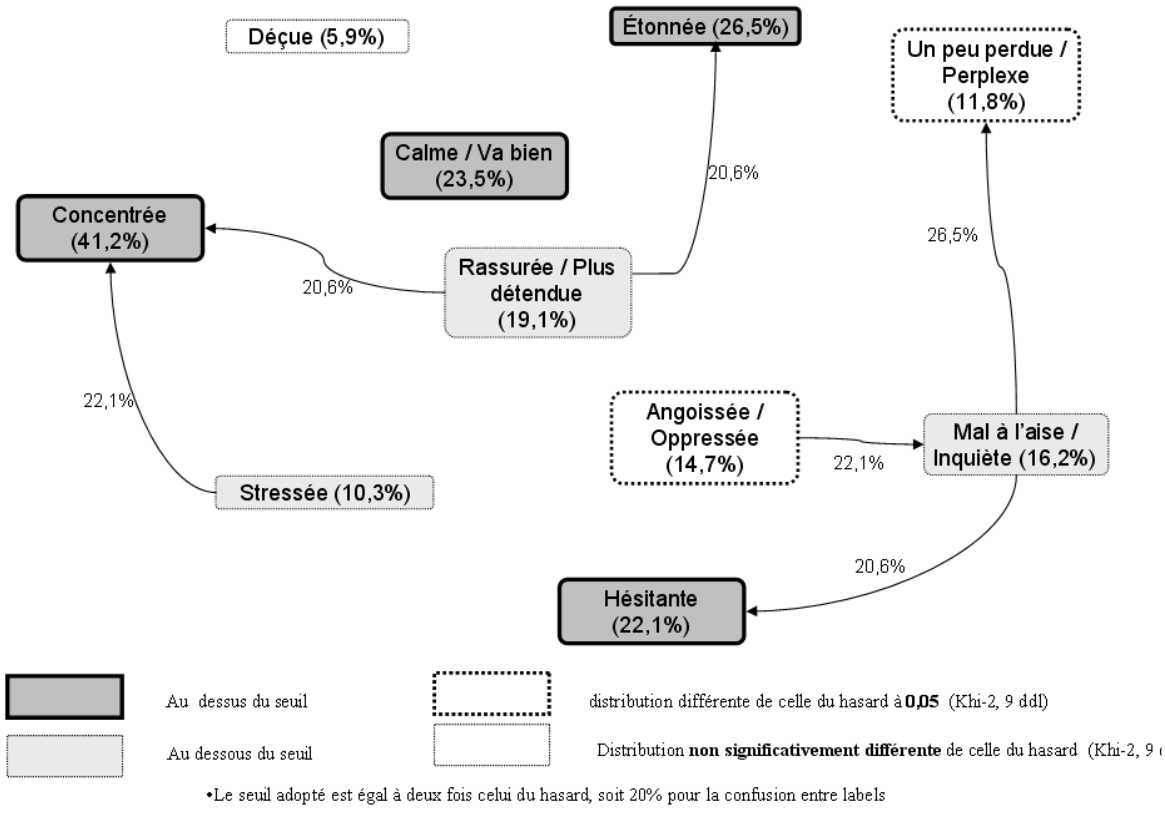



Figure 30: Résultats pour les stimuli dynamiques en condition « haut » - Sujet T
(complément de légende identique à celui de la figure 28)

La distribution des réponses pour les labels « déçue » et « un peu perdue / perplexe » en condition « haut » n'est pas significativement différente de celle du hasard (Khi-2, $p<0.05$, 9 ddl). Quant à « angoissée / oppressée », il est très bien reconnu en condition « entier » (à 36,8%) et même isolé des autres labels (donc non confondu). Pourtant, la distribution de ses réponses dans les conditions « haut » et « bas » n'est pas différente du hasard (Figures 29 et 30).

« concentrée » est, comme avec les stimuli statiques, toujours parmi les mieux reconnus (de 32,4% « bas », à 41,2% « haut ») et tend à attirer les autres réponses, et en particulier de « stressée » (dans les 3 conditions). Toutefois, lorsque les réponses aux stimuli « concentrée » se reportent sur un autre label, il s'agit de « calme / va bien » en condition « entier » et « bas ». « étonnée » est également bien reconnu dans toutes les conditions (de 26,5% en condition « haut » à 41,1% en condition « entier ».

D'autre part, « hésitante » en condition « haut », et « calme / va bien » à la fois en condition « haut » et « bas », sont curieusement mieux reconnus qu'en condition « entier ».

### III.4.2. Résultats spécifiques au sujet S.

En ce qui concerne le sujet S, en condition entier, les labels « pas concentrée et envie de rigoler », « envie de rigoler et répond au hasard » et « concentrée » ont été reconnus au dessus du seuil de deux fois le niveau du hasard (Figure 31).

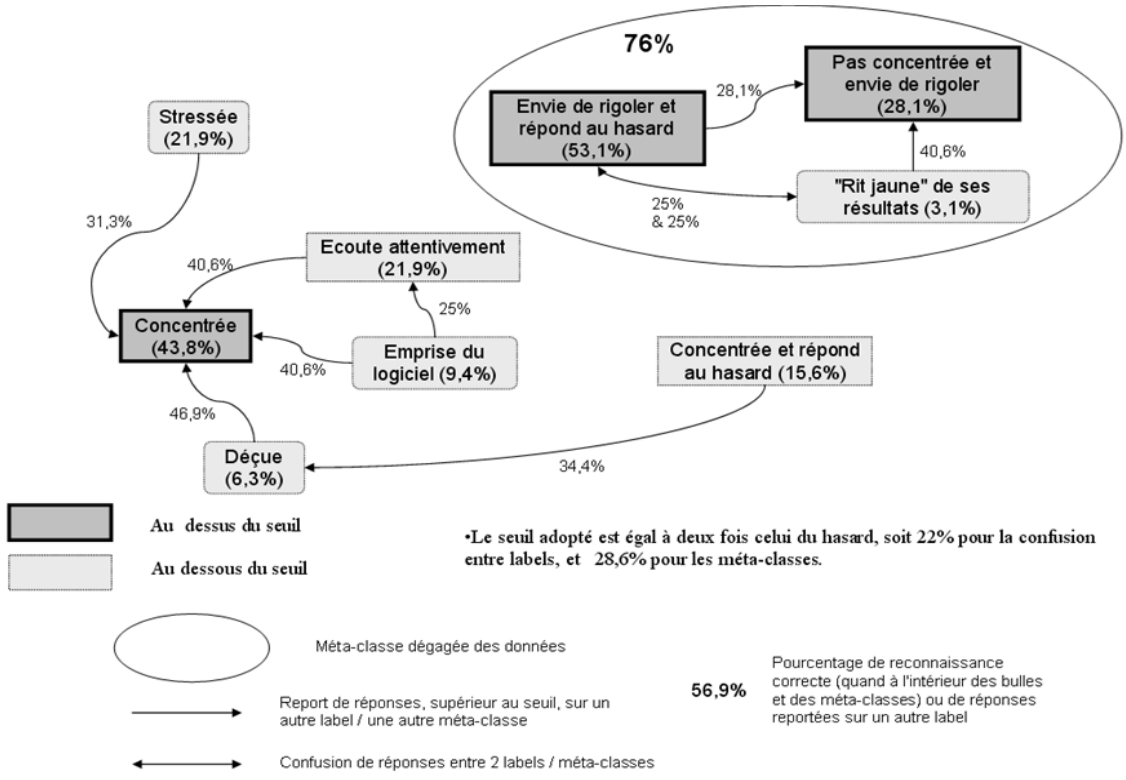



Figure 31: Résultats pour les stimuli dynamiques en condition « entier » - Sujet S

Nous avons également dégagé une méta-classe en condition « entier » : « envie de rigoler et répond au hasard » / « pas concentrée et envie de rigoler » / « "rit jaune" de ses résultats », isolée des autres labels et dans laquelle des confusions entre labels existent.

Elle a été retrouvée en condition « bas » avec des reports de réponses liés à « pas concentrée et envie de rigoler » (Figure 32), vers « stressée » pour 40,6% des réponses, et provenant de « concentrée et répond au hasard » pour 25% des réponses. Le taux d'identification global de cette méta-classe est de 62,5% en condition « bas » à 76% en condition « entier ».

Plus précisément, « écoute attentivement » est quant à lui mieux reconnu en condition « haut » qu'en condition « entier » (Figure 33) et voit la distribution de ses réponses en condition « bas » non significativement différente de celle du hasard (Khi-2, $p>0.05$, 8 ddl). Ces résultats laissent entrevoir que les informations liées à ce label se trouvent dans la partie supérieure du visage dans nos stimuli dynamiques.

La distribution des réponses pour les labels « concentrée et répond au hasard » en condition « haut », et « rit jaune de ses résultats » en condition « bas » n'est pas non plus significativement différente de celle du hasard (Khi-2, $p>0.05$, 8 ddl).

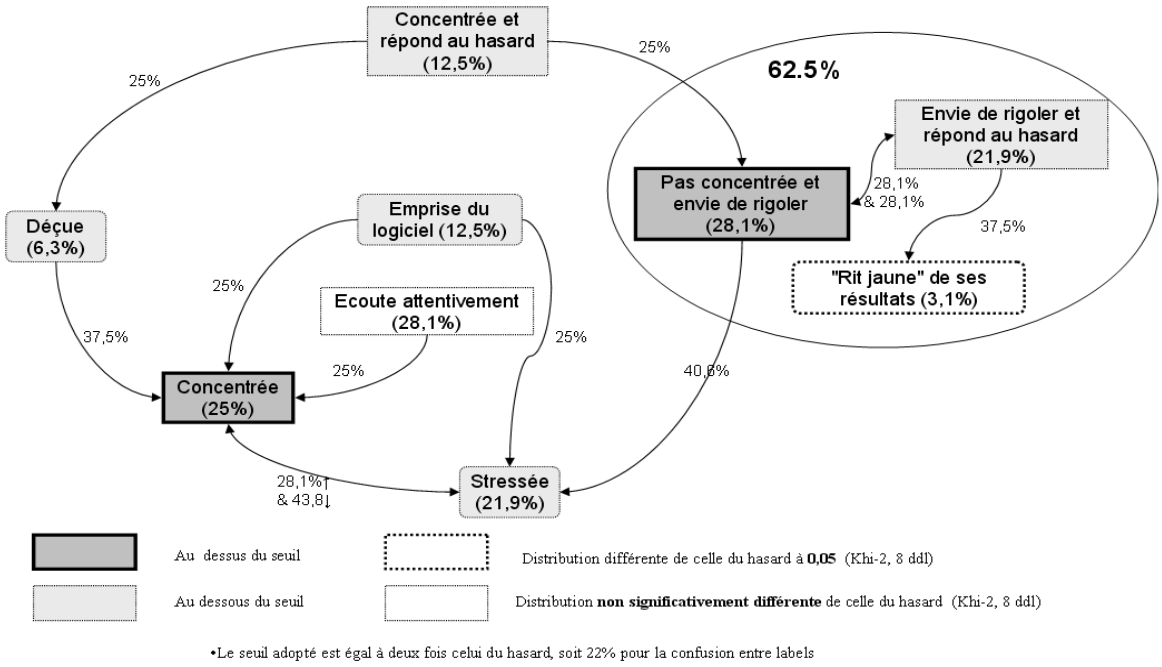



Figure 32: Résultats pour les stimuli dynamiques en condition « bas » - Sujet S
(complément de légende identique à celui du graphe de confusion précédent)

Comme pour le sujet T et comme avec les stimuli statiques, le label « concentrée », a tendance à attirer vers lui les autres réponses, et ce dans toutes les conditions (les

réponses de 4 labels différents en conditions « entier » et « bas », et celles de 5 labels en condition « haut »).

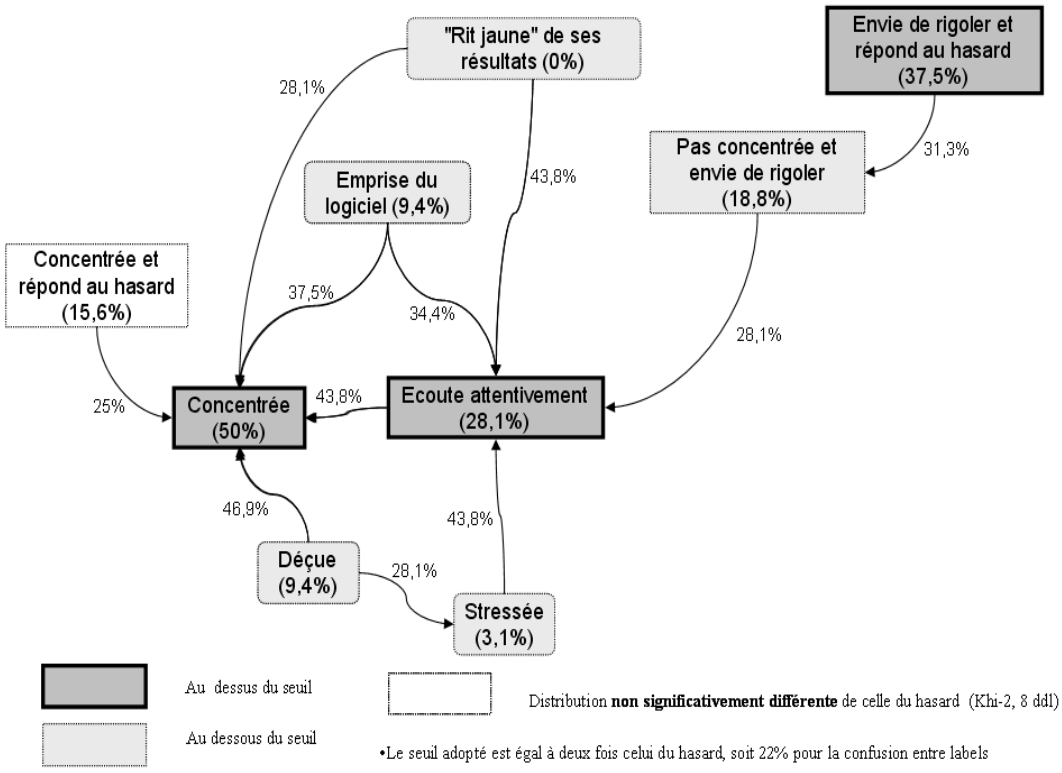



Figure 33: Résultats pour les stimuli dynamiques en condition « haut » - Sujet S (complément de légende identique à celui de la figure 31)

### III.4.3. Inter-sujets

Comme pour l'évaluation des stimuli statiques, seuls les labels « concentrée », « stressée » et « déçue » sont communs aux deux sujets et permettent ainsi une comparaison. « concentrée » est très bien reconnu pour les deux sujets et dans toutes les conditions. C'est particulièrement la condition « haut » qui obtient les taux de reconnaissance les plus élevés pour ce label, avec 41,2% pour le sujet T, et 50% pour S. De plus, il attire entre 22,1% et 31,3% les réponses de « stressée », sauf pour le sujet S en condition « haut ». « stressée » n'est par ailleurs jamais reconnus au dessus du seuil (fixé à deux fois le niveau du hasard). Quant à « déçue », la manière dont sont transmises les informations qui lui sont liées semble dépendre du sujet : il n'est bien reconnu que pour le sujet T en condition « entier ». Concernant le sujet S, 37,5% (en condition « bas ») et 46,9% (en condition « entier » et « haut » des réponses de « déçue » se reportent sur le labels « concentrée ». Il est remarquable que ce report est une tendance qui se retrouve également, mais à des taux moindres, chez le sujet T en conditions « haut » et « bas ».

## *III.5. Quelques analyses de clustering*

Nous avons, en parallèle à l'analyse des graphes de confusions, effectué des analyses de *clustering*, qui nous ont permis de valider certaines observations effectuées sur les graphes, mais aussi d'observer d'autres phénomènes. Nous présenterons seulement ici certaines des analyses répondant à un de ces deux cas.

### III.5.1. Sujet T


Figure 34: *Clustering* pour le sujet T – stimuli dynamiques en condition « entier », les phénomènes soulignés dans le texte sont entourés par les ellipses.

Selon l'analyse en *clusters* hiérarchiques pour les stimuli dynamiques du sujet T (Figure 34), en condition « entier » émerge un *cluster* « calme / va bien + rassurée / plus détendue », confirmant la méta-classe relevée (*cf.* III.4.1.).

Cette dernière se retrouve dans toutes les conditions et également en statique (mais alors couplé en conditions « haut » avec « concentrée »). Il apparaît aussi, dans cette analyse, un autre *cluster* « un peu perdue / perplexe + étonnée », et un troisième qui regroupe « stressée + mal à l'aise », « angoissée / oppressée » en étant isolé en dynamique (par rapport au statique) ce qui confirme les observations de la partie III.4.1.

Quant au *clustering* relevé en condition « haut » (Figure 35), il est globalement différent des conditions « entier » et « bas », à la fois en statique et en dynamique.

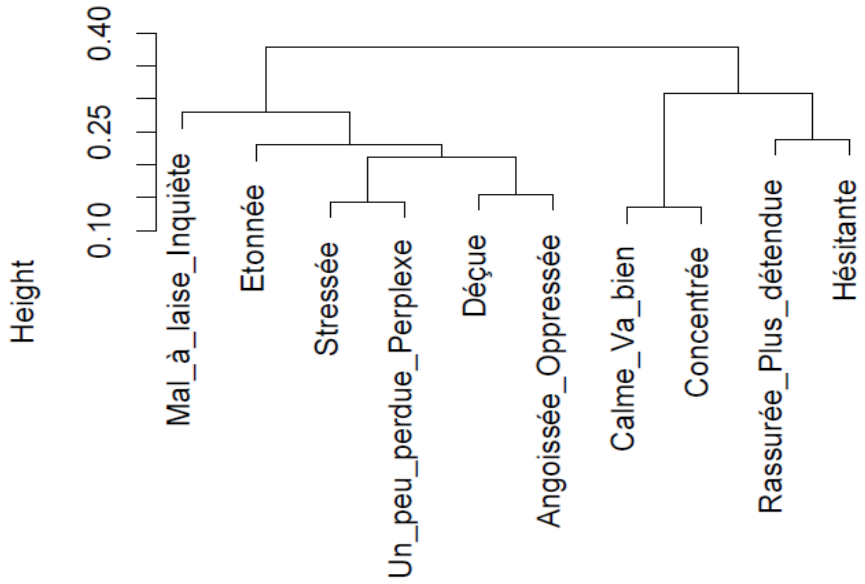



Figure 35: *Clustering* pour le sujet T – stimuli dynamiques en condition « haut »

### III.5.2. Sujet S

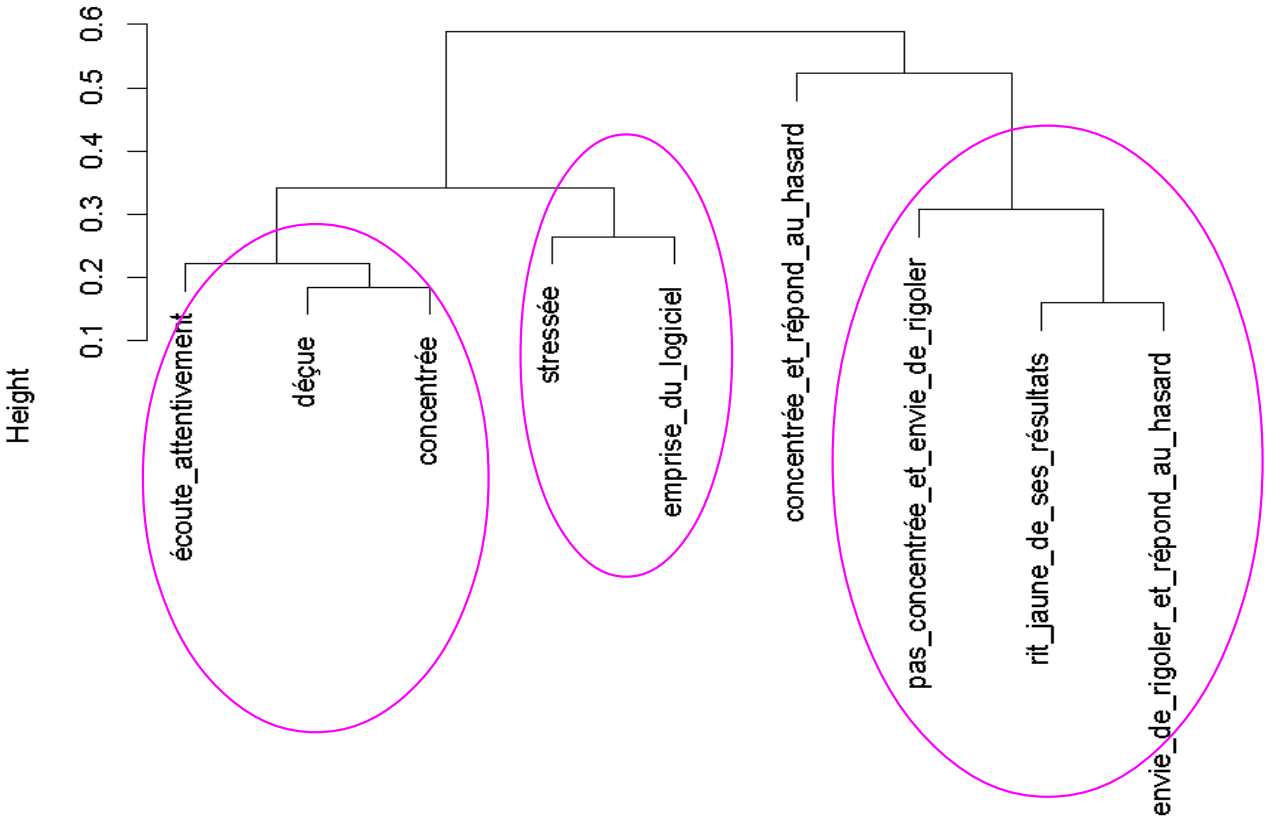



Figure 36: *Clustering* pour le sujet S – stimuli statiques en condition « entier », les phénomènes soulignés dans le texte sont entourés par les ellipses.

Concernant le sujet S, les analyses avec les stimuli statiques et en condition « entier » (Figure 36) indique une organisation des labels en plusieurs groupes :

- d'une part les trois labels incluant la notion de « rire », correspondant à la méta-classe dégagée, rattachés au label « concentrée et répond au hasard » qui partage en partie avec ce *cluster* la notion de « réponse au hasard » (en particulier par « envie de rigoler et répond au hasard ») ;
- d'autre part « écoute attentivement + déçue + concentrée » et « stressée + emprise du logiciel ».

En ce qui concerne les stimuli dynamique, l'analyse en *clustering* (Figure 37) nous indique deux gros blocs : d'un côté « envie de rigoler et répond au hasard + pas concentrée et envie de rigoler + rit jaune de ses résultats » avec « concentrée et répond au hasard » dans ce même groupe mais un peu à l'écart ; de l'autre un bloc de deux *clusters* : « concentrée + déçue » (retrouvé en condition bas) et « écoute attentivement + emprise du logiciel + stressée ».

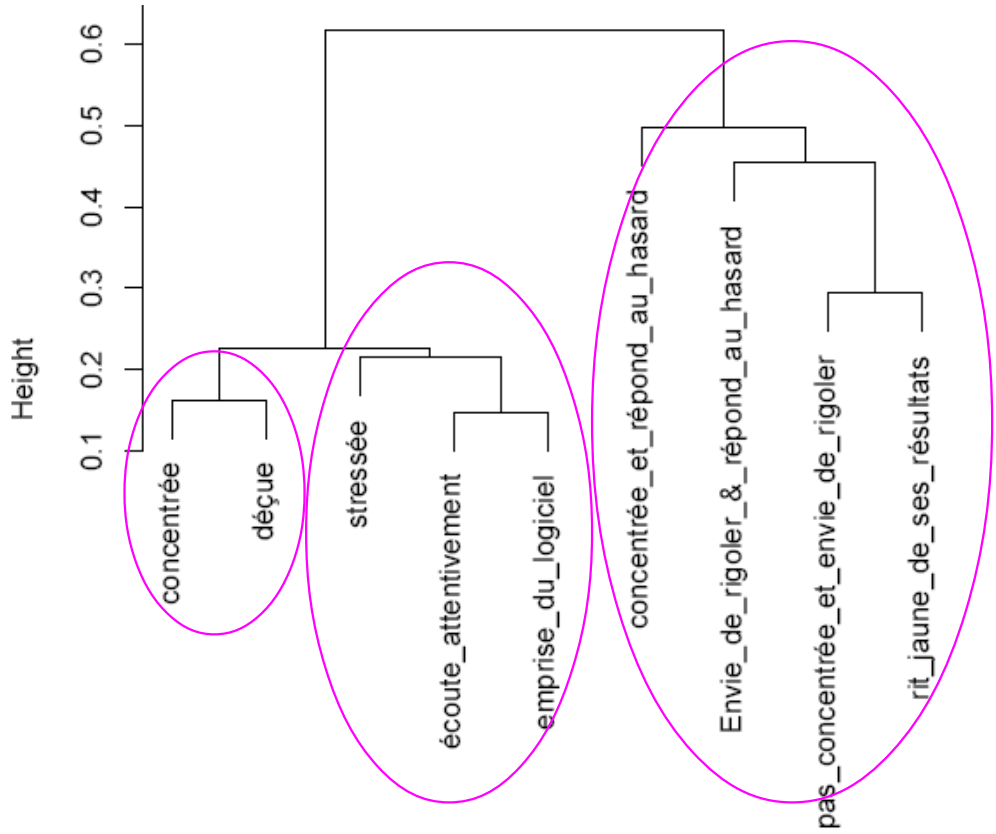



Figure 37: *Clustering* pour le sujet S – stimuli dynamiques en condition « entier », les phénomènes soulignés dans le texte sont entourés par les ellipses.

En condition « bas », un *cluster* restreint et concernant les deux labels contenant l'idée de réponse au hasard apparaît (Figure 38) : « envie de rigoler et répond au hasard + concentrée et répond au hasard ».

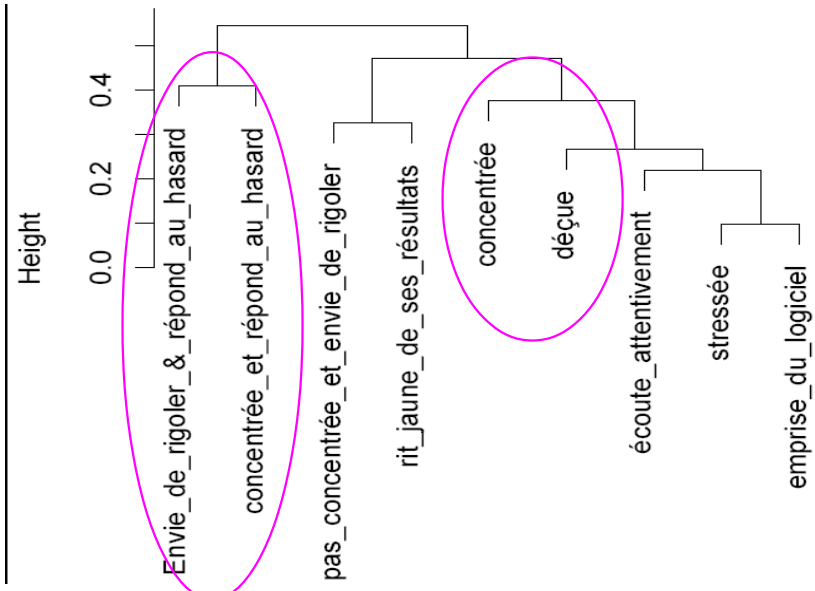



Figure 38: *Clustering* pour le sujet S – stimuli dynamiques en condition « bas », les phénomènes soulignés dans le texte sont entourés par les ellipses.

## *III.6. Interaction statique / dynamique*

Les résultats obtenus lors du test des stimuli dynamiques ont été différents de ceux du test des stimuli statiques.

### III.6.1. Résultats spécifiques au sujet T.

Les quatre méta-classes identifiées lors du test statique (« hésitante » / « stressée » / « mal à l'aise / inquiète » / « angoissée / oppressée » ; « un peu perdue / perplexe » / « déçue » / « étonnée » ; « rassurée / plus détendue » / « calme / va bien » ; et « concentrée », n'ont pas été retrouvées de manière identique pour les mêmes stimuli dynamiques. En effet, si les deux dernières ont été retrouvées telles qu'elles, « angoissée / oppressée » s'est isolé en dynamique et les autres labels n'ont formé plus qu'un seul groupe.

Nous observons donc (Figure 22 et Figure 28) :

- une fusion de « hésitante », « mal à l'aise / inquiète », « stressée » et « angoissée / oppressée » d'une part, et « étonnée », « déçue », « un peu perdue / perplexe » d'autre part, lors du passage du statique au dynamique ;
- l'isolement de « angoissée / oppressée » en dynamique.

Les labels « étonnée » dans toutes les conditions, « angoissée / oppressée » en condition « entier », et « hésitante » en condition « haut », ont été mieux reconnus en dynamique, alors que « un peu perdue / perplexe » en condition « entier », « angoissée / oppressée » en condition « haut », et « déçue » en condition « bas » l'ont mieux été en statique.

### III.6.2. Résultats spécifiques au sujet S.

En dynamique et en condition « entier », tous les labels ont eu une distribution de leurs réponses différentes de celle du hasard (Khi-2, $p<0.01$, 8ddl), alors qu'en statique, les réponses aux stimuli « pas concentrée et envie de rigoler » étaient données au hasard.

De plus, alors que « pas concentrée et envie de rigoler » en conditions « entier » et « bas », et « concentrée » en conditions « haut » et « bas » ont été mieux reconnus en dynamique, « écoute attentivement » l'a mieux été en statique pour la condition « entier » (Figure 25 et Figure 31).

Par ailleurs, la méta-classe identifiée en dynamique est restée la même que celle relevée en statique.

# IV. Discussion des résultats et perspectives

## *IV.1. Synthèse des résultats : une perception des expressions où la globalité comme les « détails » sont pertinents*

Un des résultats les plus intéressants, si l'on se réfère à des modèles combinatoires des expressions faciales (comme le FACS), est qu'il ne semble pas, d'après nos résultats pour ces expressions, que les informations véhiculées par le haut et par le bas du visage soient purement additives ou complétives en rapport des informations véhiculées par la face en entier. Par ailleurs, aucune des deux parties du visage ne contient une information nulle, cela quelque soit le label, et en particulier en ce qui concerne l'expression des états mentaux. L'additivité de type « zéro + tout », ne semble pas non plus être valide, même en utilisant uniquement la dichotomie, positive vs. négative, de la valence des affects (le bas du visage ne suffit pas à la valence positive et porte également la valence négative, et inversement pour le haut).

Toutefois, certains labels ont plus d'informations dans le haut ou le bas du visage, et cette répartition peut changer en fonction de la nature dynamique ou statique du stimulus présenté. Ainsi, quand nous passons du statique au dynamique :

- les informations liées à « concentrée », « envie de rigoler et répond au hasard » se concentrent dans le haut du visage, comme celles d'« hésitante », ou celles d'« écoute attentivement » qui y restent ;
- les informations liées à « stressée » passent du haut au bas du visage, comme celles de « calme / va bien » ou « pas concentrée et envie de rigoler », et de « rassurée / plus détendue » qui y restent ;
- les informations liées à « déçue » et « angoissée / oppressée », auparavant respectivement dans le bas et le haut du visage, se retrouvent alors dans le visage dans son ensemble, comme pour « étonnée ».

Par ailleurs, les résultats obtenus pour les stimuli dynamiques ont été différents de ceux du test sur les stimuli statiques. Nous noterons en particulier que certains labels ont été mieux identifiés avec les stimuli de nature statique qu'avec ceux de nature dynamique. Il s'agit notamment :

- pour le sujet T (Figure 39), de « un peu perdue / perplexe » en condition « entier », « concentrée » et « déçue » en condition « bas », et « angoissée / oppressée » en condition « haut » (mieux reconnu en dynamique en condition « entier »).

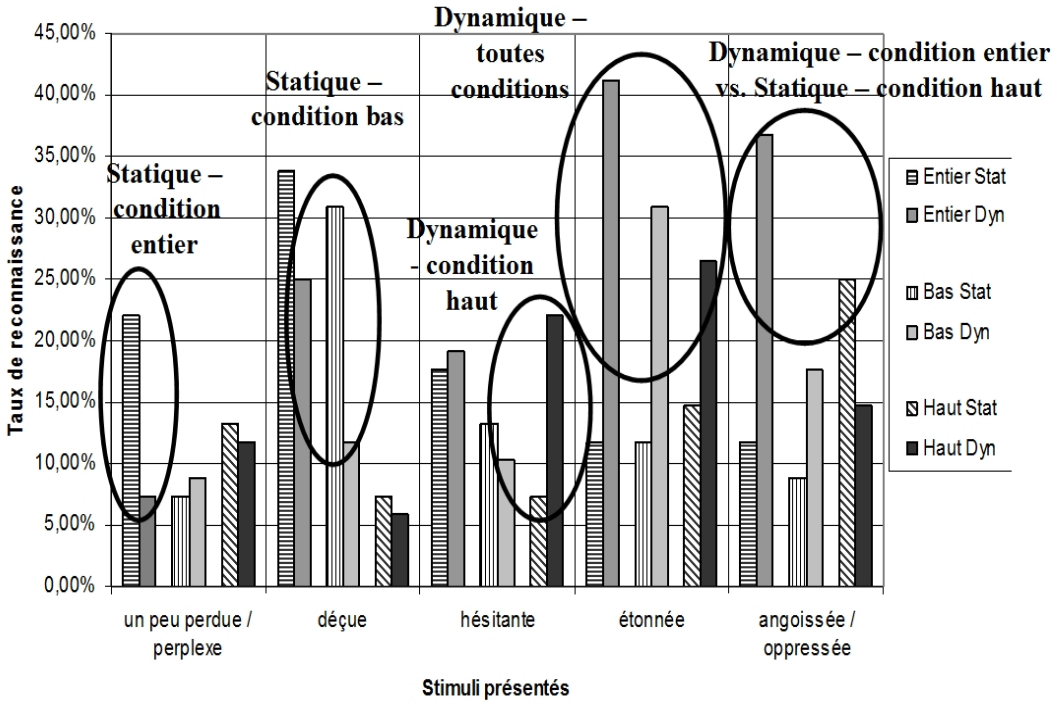



Figure 39: Comparaison statique / dynamique pour quelques labels – Sujet T.
Les ellipses et leur description focalisent sur les conditions les mieux reconnues pour ces labels.

- pour le sujet S (Figure 40), de « écoute attentivement » en conditions « entier » et « haut », et « déçue » en condition « haut ».

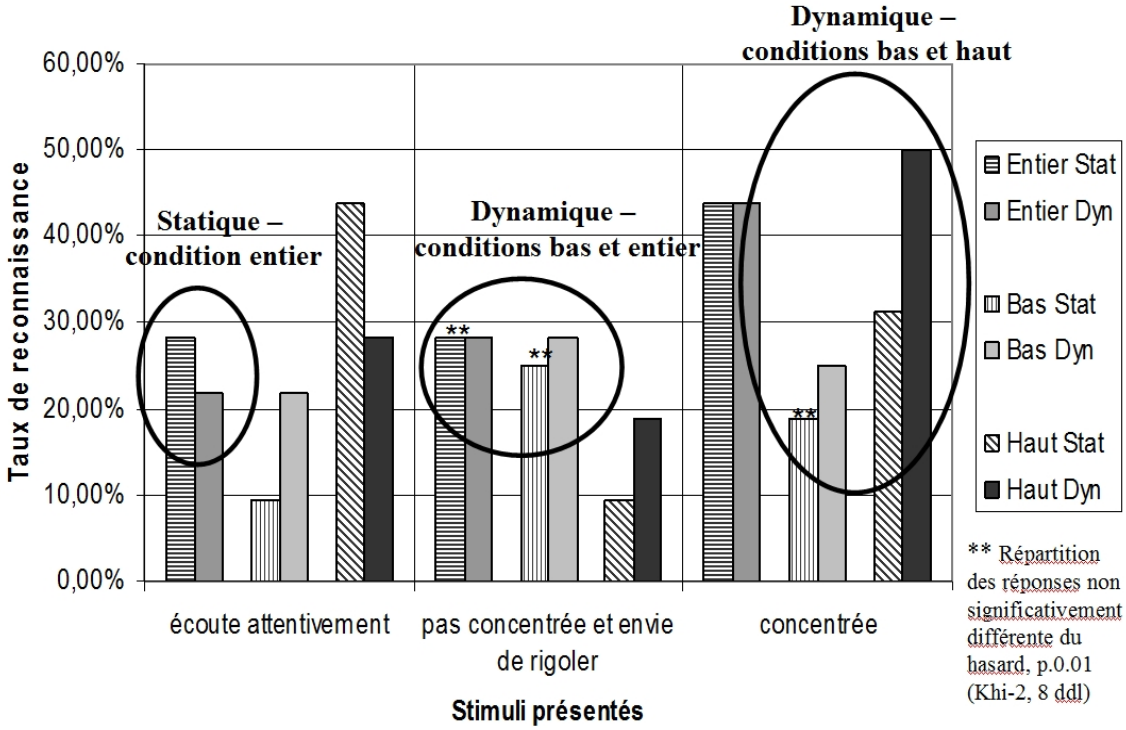



Figure 40: Comparaison statique / dynamique pour quelques labels – Sujet S.
Les ellipses et leur description focalisent sur les conditions les mieux reconnues pour ces labels.

Globalement, le gain d'information lors du passage du statique au dynamique semble dépendre de manière importante de la nature de l'information : pour certains stimuli, un taux de reconnaissance en dessous du seuil du hasard peut devenir clairement au dessus de ce seuil, alors que pour d'autres, le dynamique semble perturber. Peut-être pouvons revenir sur un point méthodologique : lorsqu'un stimulus statique est présenté, il l'est pendant un certain temps par nécessité méthodologique (temps que nous avons laissé à la liberté du sujet, comme classiquement dans beaucoup d'expériences en visuel statique). Cela introduit bien entendu une ambiguïté quant à la nature du processus testé. En effet, dans les paramètres qui caractérisent les stimuli dynamiques (testés avec un temps imposé limite), le paramètre du « temps d'exposition » peut être le seul paramètre « dynamique » pertinent, en particulier lorsque le visage est immobile. La durée d'immobilité est alors le facteur informatif dans le processus de reconnaissance dynamique.

De la même manière, lors d'une tâche de présentation de stimuli dynamiques, le sujet peut, sans que nous n'en ayons la trace, mettre en œuvre un processus de reconnaissance d'images statiques qu'il récupère dans le stimulus dynamique. Inversement dans une tâche de présentation de stimuli statiques, le temps d'exposition des images peut devenir un artefact utilisé en reconnaissance dynamique de traces chimères.

Afin d'approfondir ces hypothèses, et de préciser quels sont les paramètres pertinents d'un point de vue communicatif, il serait intéressant :

1) de relever quels sont les types de mouvements qui ont été les mieux reconnus en statique et en dynamique ;
2) de chercher dans nos vidéos des mouvements similaires, mais avec des paramètres différents (comme l'amplitude, la répétition, la furtivité) ;
3) de mettre en place un test perceptif inverse de celui effectué ici, c'est-à-dire en donnant à percevoir ces mouvements particuliers, mais en laissant les juges naïfs décider quel état est exprimé.

## *IV.2. Des modèles de gestualité et d'expressions faciales des affects incomplets*

Les modèles que nous avons évoqués au cours des précédents chapitres apportent, la plupart du temps, des méthodologies et des grilles d'annotation adaptées à ce pour quoi ils sont faits. Cependant il apparaît qu'ils ne permettent pas toujours d'étiqueter ou d'expliquer les icônes gestuelles et les phénomènes perceptifs que nous observons dans nos données et à travers nos résultats.

Sans revenir sur les considérations qui nous ont incités à nous affranchir de l'influence de ces modèles (*cf.* entre autres Chapitre 3 II.3.), nous allons seulement mentionner les limites que nous avons observées concernant le modèle le plus largement utilisé : le FACS d'Ekman (*cf.* Chapitre 2 I.2.2. et p.109).

Outre le fait que les AUs ne permettent pas d'étiqueter et de modéliser certaines de nos icônes gestuelles (*cf.* Chapitre 3 IV.4.), nous avons montré que l'additivité entre le haut et le bas du visage en termes d'information ne se vérifiait pas sur nos données. Cela implique également que le FACS n'est pas adapté à la modélisation des expressions d'« émotions mélangées » (*blended emotions*), et des expressions de d'autres mélanges d'états (affectifs ou cognitifs) comme ceux désignés par les auto-annotations de nos sujets (*cf.* Chapitre 2 I.3. et partie II.2. de ce chapitre).

Par ailleurs, Ekman propose dans son hypothèse catégorielle que chaque émotion est associée à une expression prototypique, universelle pour ses émotions de base. Pourtant, nous avons proposé à la perception deux ou quatre icônes gestuelles (selon le sujet) pour chacun des labels. Ces IG étaient différentes et choisies de telle façon que les postures et mouvements décrits soient a priori bien distincts (Figure 41). Or, certains labels ont été très bien reconnus globalement, c'est-à-dire que les juges ont associé chacune des icônes testées au label correspondant.


Figure 41: Exemple d'IGs du sujet T, auto-annotées toutes deux par le label « déçue », bien reconnu lors des tests perceptifs

Ainsi, plusieurs expressions faciales (nos icônes gestuelles) peuvent correspondre à un seul et même label. Cela signifie soit que notre description d'IGs correspond à un degré de finesse qui distingue des variantes d'un unique prototype pour les deux ou quatre IGs[70], soit qu'il s'agit de prototypes différents.

Quant à l'existence de *display rules,* liées au contexte social de l'interaction, qui réguleraient les expressions (*cf.* Chapitre 2 II.3.1.), nous avons pu constater la présence d'attitudes, c'est-à-dire d'expressions sociales, dans nos données d'IHM (donc sans la présence d'un autre humain). Ainsi, le sujet régulerait et contrôlerait les expressions en dehors d'un contexte social tel qu'il est envisagé par Ekman. L'influence du contexte sur les expressions doit donc être pris dans un sens beaucoup plus large.

Enfin, nos résultats vont dans le même sens que ceux de Kaiser, Wherle & Schmidt (1998, *cf.* la partie I.4. de ce chapitre), concernant la critique du FACS, quant à sa modélisation de la dynamique, qui ne semble pas adaptée. En effet, nos résultats nous amènent à faire l'hypothèse que c'est dans certains cas le mouvement lui-même et sa dynamique qui est informatif de l'état du sujet. Or, le FACS ne permet pas de modéliser ou de générer précisément la dynamique des expressions faciales.

## *IV.3. Quelques perspectives*

### IV.3.1. Indices temporels et dynamique du mouvement

Nous avons suggéré dans la partie IV.1. l'importance d'indices temporels (tels que le temps d'exposition des stimuli ou la durée d'un geste) dans l'interprétation d'expressions faciales, mais aussi l'importance de la dynamique des mouvements (partie précédente).

Nos données nous amènent à concevoir la dynamique du mouvement en termes de vitesse, mais surtout en termes d'accélération. En effet, nous avons relevé lors de l'étiquetage des icônes gestuelles, qui seraient qualifiées de « furtives » dans la langue courante. Ces icônes concernent la plupart du temps les lèvres (IGLe), les yeux (IGY) ou les sourcils (IGSour). Mais quels sont les paramètres qui font que nous les caractérisons de « furtives » ? Qu'est ce qui fait que nous les distinguons d'un mouvement identique et de même amplitude, mais « non furtif ». Les expressions de ce type transmettent-elles une information différente de celle d'une icône « normale » de même type, et si oui à quel niveau ?

Les premiers éléments de réponses à ces questions nous viennent de la notion d'amorce de mouvement : il semblerait que ces icônes « furtives » soit caractérisées par une amorce de mouvement, avec son importante accélération, mais sans le

[70] Dans ce cas la configuration commune prototypique n'est pas du tout évidente

mouvement qui suit. Il serait possible d'approfondir cette problématique en testant la perception de ce type d'icône sous forme de *gating* (c'est-à-dire en donnant à percevoir d'abord les premières millisecondes du mouvement, puis en allongeant ce temps progressivement, jusqu'au mouvement complet - *cf.* Audibert, Aubergé, & Rilliard, 2007 -). La perception pourrait être testée en termes d'information transmise, et / ou de perception catégorielle ou non (au sein de cette gradation de durée). Une telle expérimentation nous permettrait aussi de vérifier la pertinence communicative du paramètre d'accélération (*vs.* vitesse seule).

Ainsi, notre hypothèse est que la dynamique du mouvement est perçue en tant qu'information qui nous indique la « cible » de ce dernier.

D'autres éléments de réponses pourraient nous venir des théories du mouvement biologique (*e.g.* Viviani & Stucchi, 1989 ; 1992 ; ou plus récemment Gibet, Kamp, & Poirier, 2004). En effet, ces dernières ont entre autres montré que selon la dynamique d'une trajectoire (c'est-à-dire la vitesse et l'accélération) qui simule le dessin d'un cercle sur un écran, la perception va être celle d'un cercle ou alors d'une ellipse. Ce type de recherches montre ainsi l'importance de la dynamique dans la perception que nous avons du mouvement.

### IV.3.2. Multimodalité et micro-événements audibles

Nos recherches ont porté jusqu'alors sur des événements appartenant à la modalité visuelle. Or, d'un point de vue écologique, ces événements sont perçus de manière audio-visuelle : certains mouvements sont également audibles. En parallèle, les micro-événements audibles, tels les interjections et autres « bruits de bouche », s'ancrent nécessairement dans un mouvement, visible ou non. Le son / le bruit produit est lié au mouvement sous-jacents.

De plus, un premier travail de Loyau sur les rires de notre corpus (Loyau, 2007) a montré que les sujets sont capables de replacer un rire dans un énoncé, en terme d'organisation (*e.g.* choisir si le rire apparaît au début, milieu ou fin d'énoncé). Ainsi, un rire pourrait soit indiquer le contexte, soit être intrinsèquement contextuel. Qu'en est-il alors des interjections et autres événements vocaux ?

À côté de ces observations, les travaux antérieurs portant sur certains types particuliers d'événements vocaux (voir Chapitre 5 I. et II.) confirment l'idée que ces micro-événements audibles de nature variée auraient une pertinence communicationnelle. Cela nous a amenés à poursuivre notre étude de l'expression du *Feeling of Thinking* en portant particulièrement notre attention sur ces événements audibles de notre corpus.

Il semblerait que ces différents événements vocaux soient produits de manière variable à la fois, au niveau inter-sujets, dans le choix de leur type et dans leur utilisation, et au niveau intra-individuel, selon la tâche effectuée par le sujet et de son état mental et affectif. De plus, leur organisation temporelle semble loin d'être aléatoire. Encore plus que pour les expressions gestuelle/faciales, cette variabilité dans la manière d'utiliser les événements vocaux nous a amené à augmenter le nombre de données vocales sur lesquelles travailler. Nous avons pour cela étendu l'étiquetage puis l'analyse des événements vocaux à quatre autres sujets (désormais, en tout, trois hommes et trois femmes, dont le choix est expliqué Chapitre 5 III.1.3.). Nous avons ensuite cherché à comparer, c'est-à-dire trouver des régularités et des différences, entre le comportement des différents sujets.

### IV.3.3. Pertinence de la rythmicité

À moyen terme, il serait ensuite intéressant de chercher à vérifier, pour le *Feeling of Thinking,* si la rythmicité d'un geste (comme sa régularité et sa fréquence) est un indice fort de l'état affectif dans lequel le sujet étudié se trouve. En effet, Carlier et Graff ont montré que chez les joueurs de tennis de haut niveau, en situation de match, un événement ponctuel, un comportement particulier au joueur, du type « amener la main tenant la raquette à la taille », était d'autant plus fréquent et irrégulier que le joueur était stressé, et cela était perçu naïvement par le public (Carlier & Graff, 2006). Il nous reste donc à tester le rôle de la rythmicité de certaines de nos icônes gestuelles et vocales dans une tâche d'interaction. Cependant, nous sommes confronté à la difficulté de ne pas savoir à quelle granularité (*e.g.* mouvement dans la partie haute du visage, mouvement quelconque du sourcil, uniquement haussement de sourcil, avec éventuellement seulement tel ou tel paramètre) et à quel niveau temporel analyser cette rythmicité, d'autant plus que les indices visuels dans la communication expressive sont continus dans le temps.

Cela nous amène ainsi à nous focaliser dans un premier temps sur la temporalité, la rythmicité des événements vocaux, ceux-ci étant des événements ponctuels dans le temps, ce qui facilite la tâche par rapport à l'analyse de l'organisation temporelle des événements visibles.

# V. Résumé

Nous venons de voir que certaines Icônes Gestuelles ont été testées perceptivement afin de les valider. Pour étudier dans quelle mesure l'information de *FoT* est portée par le statique ou la dynamique, deux tests perceptifs identiques ont été implémentés : le premier avec des formes statiques des stimuli (photos), le second avec les mêmes Icônes Gestuelles sous leur forme dynamique (vidéos). De plus, afin d'évaluer l'hypothèse de certains travaux autour de la théorie d'Ekman quant à la répartition de l'information statique sur différentes parties du visage, trois conditions de présentation ont été testées : visage entier et haut et bas du visage seuls. Les icônes ont été testés dans une tâche d'association entre stimuli et labels d'auto-annotation qui leur étaient attribuées par les sujets eux-mêmes (à propos de leurs états mentaux / émotionnels).

Les résultats montrent que certaines Icônes Gestuelles ont été mieux reconnues avec le stimulus statique que le stimulus dynamique correspondant. Globalement, le bénéfice d'information du statique au dynamique semble profondément dépendre de la nature de l'information, puisque la dynamique peut perturber l'identification de label. Ainsi, alors que certains événements visuels sont décodées en utilisant un processus de perception statique, avec néanmoins une durée de présentation des stimuli probablement significative, d'autres gestes sont en parallèle décodés dynamiquement.

Nous avons aussi montré que les informations visuelles concernant le haut et le bas du visage ne sont pas additive, quelque soit la nature ou la valence de l'état exprimé. De plus, l'amplitude des mouvements de la face ne peut être directement considérée comme un paramètre et doit être reliée au comportement du locuteur : *e.g.* les mouvements d'un sujet introverti peuvent être très informatifs sans avoir une grande amplitude.

Par ailleurs, il semblerait que l'information de *FoT* soit portée dans certains cas par des motifs temporels globaux. Cette hypothèse porte sur certains gestes ou événements vocaux souvent idiosyncrasiques, sans valeur d'états affectifs ou mentaux intrinsèque, mais dont ces derniers pourraient être révélés à travers leur organisation temporelle, leur rythmicité au cours de l'interaction. De tels événements concerneraient alors le comportement global du locuteur plutôt que l'observation d'événements ponctuels, et va dans le sens des résultats de Carlier & Graff (2006) dans leurs travaux concernant le comportement de joueurs de tennis de haut niveau.

En parallèle, l'étiquetage du corpus a révélé une grande quantité d'événements vocaux, *a priori* dépendant en particulier du comportement du locuteur. Les questions sont alors :

- quelle est la nature de ces événements vocaux?
- quelle type d'information est portée par la modalité auditive en dehors des tours de parole du sujet?
- dans quelle mesure ces événements vocaux non-lexicaux ou pré-lexicaux apporte de l'information sur les états de *FoT* des sujets?

# CHAPITRE 5 : LES MOUVEMENTS AUDIBLES DU CONDUIT VOCAL : DE LA THÉORIE À LA PRATIQUE

Dans le chapitre précédent, nous avons montré la pertinence de la modélisation des expressions du *FoT* à travers les modalités visuelles de la gestualité et des expressions faciales.

Depuis le début des études sur l'interaction humaine, en particulier dans le domaine de la pragmatique, fortement influencée par la théorie des actes du langage de Austin (1975) puis Searle (1979), ce qui constitue le tour de parole, ce qui est langage parlé, se distingue de ce qui se situe hors du tour de parole. Nous considérons plutôt ici la communication comme une continuité communicative, ponctuée par les événements que sont les prises de parole (Aubergé 2002a). Le canal acoustique devient évidemment mineur quand le locuteur suit le tour de parole de son interlocuteur, ou quand il prépare sa réponse à la tâche en IHM. Toutefois, et nous le montrerons dans la suite de ce chapitre, ce temps en dehors de la parole est ponctué d'événements du *FoT,* eux-mêmes suivis par l'interlocuteur parlant. Le canal visuel varie en permanence, à travers des mouvements et postures faciaux et gestuels, dont nous avons montré également la pertinence à véhiculer des valeurs du *FoT* au cours du Chapitre 4.

Rappelons que nos sujets ne sont pas en situation d'écoute d'un interlocuteur humain, mais en situation de lecture, ou d'écoute avec une tâche a résoudre. Les indices de *FoT* produits par les sujets sont donc relatifs à cette tâche d'IHM. Toutefois, nous montrerons que des sujets humains les perçoivent et les « comprennent ». Cela tend à montrer que les indices que nous allons mettre en évidence ne sont pas spécifiques à l'IHM, mais bien propres à la communication humaine. Cependant, l'occurrence de ces indices de *FoT* est sans doute à relier à une telle situation IHM.

Pour les six sujets étudiés dans cette partie de nos recherches, la proportion de parole dans l'interaction globale est de l'ordre de 17 à 25% (voir analyses plus complètes chapitre 6).

Nous relevons des événements sonores audibles, de nature variée, produits par le conduit vocal, du « bruit de bouche » à des objets pré-lexicalisés comme les interjections. C'est en parlant de ces éléments (du moins de la partie correspondant aux interjections), qu'Ameka a introduit la notion de geste vocal (*vocal gesture)*, avec deux idées sous-jacentes (Ameka, 1992) :

- qu'il existe une relation étroite entre interjections et gestes en général ;
- que les interjections sont à la frontière entre communication verbale et non-verbale.

Ainsi, comme l'a souligné Wilkins (1992) avec un point de vue linguistique, c'est l'étude des périphéries du langage qui nous apportera les réponses à beaucoup des préoccupations linguistiques actuelles.

C'est donc dans cette perspective à la fois multimodale, linguistique, pragmatique, mais également dans les applications technologiques, que nous allons ancrer notre étude de ces événements, de ces gestes vocaux.

# I. Pourquoi et dans quel but étudier les événements vocaux non-lexicaux

## *I.1. Enjeux théoriques*

### I.1.1. Multimodalité et notion de « geste vocal »

De la même manière qu'il n'est pas écologique de percevoir une image statique, une vidéo sans signal audio n'est pas complète d'un point de vue écologique. D'autant plus que l'importance du non-verbal et la multimodalité de la communication sont communément acceptées (Mehrabian & Wiener, 1967).

Comme nous l'avons évoqué Chapitre 2 I.1.2., cette multimodalité peut désigner deux phénomènes différents. C'est à la multimodalité de type « intrinsèque », à la micro-bimodalité de Poggi (2007) que se rapporte la notion de « geste vocal » (du côté de la production).

Cette notion de « geste vocal » (*vocal gesture*) a été introduite par (Ameka, 1992) avec les idées que les interjections sont à la frontière entre communication verbale et non-verbale, mais aussi qu'il existe une relation étroite entre la forme audible des interjections et les gestes en général. Ce lien est particulièrement mis en évidence par les recherches d'Eastman (1992) sur les interjections (d'un point de vue fonctionnel) en Swahili. Cette dernière a montré que les interjections de cette langue peuvent être des expressions verbales seules ou accompagnées de gestes, ou encore des gestes seuls à fonction interjectionnelle[71]. Ainsi, le cas du swahili met en évidence que l'utilisation du langage et du geste « implique des formes pragmatiques de communication qui ne peuvent être séparées par des frontières déterminées. Les interjections apparaissent ainsi où ces frontières sont plutôt floues »[72] (*ibid*, p.285).

De plus, et comme c'est le cas pour des événements vocaux que nous relevons dans notre corpus, « certaines formes avec une fonction émotionnelle ne peuvent être incluses dans les dictionnaires, soit parce qu'elles sont complètement impliquées par le mouvement, soit parce qu'elles ne sont pas en conformité avec la structure phonologique de la langue décrite ».[73] (*ibid*, p.275)

[71] Citation originale : « Interjections in Swahili, seen this way, range from verbal expressions alone, to verbal expressions accompanied by gesture, to gestures alone used with an interjectional function. »

[72] Citation originale : « […] the evidence from Swahili is that language-use and gesture-use, in part, involve pragmatic forms of communication not separable by any definite boundaries. Interjections occur where boundaries are rather fuzzy. »

[73] Citation originale : « Some forms with interjectional function defy dictionary inclusion being totally movement-involved or out of conformity with the language's phonological structure. »

Scherer (1994, p.179 et 182) a souligné la multimodalité des *affects bursts,* dans le sens où gestes articulatoires et sons produits sont étroitement liés. Il les a ainsi qualifiés « d'éléments particuliers du comportement, [...] d'exemple primordial d'expression qui intègre facial et vocal »[74] (*ibid*, abstract).

C'est pourquoi même dans le domaine de l'*Affective Computing,* en Interaction Homme-Machine, modéliser la multimodalité (à la fois « globale » et « intrinsèque ») de la communication, est l'enjeu d'une part de plus en plus importante des recherches (Pelachaud & Poggi, 2002).

### I.1.2. Des événements vocaux nombreux et aux paramètres informationnels variés

Pour notre part, et y compris en dehors des tours de parole, nous relevons un grand nombre d'événements vocaux dans notre corpus, ce qui confirme et étend l'observation de Ward (2000) qui, lors de conversations en anglais américain, relève en moyenne une occurrence toute les 5 secondes de ce qu'il nomme *conversational grunts* (qui désigne une catégorie plus restreinte et incluse dans nos événements vocaux, voir plus loin). Campbell (2007, p.119) a également confirmé dans une étude, à la suite d'Alwood (1995), « qu'une grande quantité (presque la moitié) des sons de parole utilisés dans une conversation sont non-verbaux, souvent perçus comme du "bruit", mais qui ont pour fonction d'indiquer d'importantes informations liées aux affects »[75].

Ils sont donc nombreux, mais pourtant non traités la plupart du temps. Il s'agit en quelque sorte d'une « classe-poubelle » :

> « Pourtant, comme beaucoup de ces sons de parole non-verbaux sont typiquement considérés comme des *fillers* [littéralement "remplisseurs"], des "hésitations", des "erreurs de performance discursive" (sic), ou comme un manque évident de préparation du discours, ils sont fréquemment supprimés des enregistrements, non annotés dans les transcriptions, ou simplement non produits par les locuteurs professionnels (acteurs, publicitaires, journalistes, etc.), à qui de nombreux chercheurs font confiance pour produire leurs données d'analyse. »[76] (Campbell, 2007, p.119)

[74] Citation originale : « I then introduce a particular piece of behavior, which I shall call "affect burst", as a prime example of integrated facial vocal expression. »

[75] Citation originale : « that a large amount, approximately half, of the speech sounds used in normal everyday conversational speech are nonverbal, often simply perceived as `noise' but functioning to signal important affect-related information. »

[76] Citation originale : « However, because many of these nonverbal speech sounds are typically considered as fillers. or .hesitations., .performance errors. (sic), or as evidence of lack of preparation of the speech utterance, they are frequently edited out of recordings, disregarded in a transcription, or simply not produced at all by the professional speakers (actors, announcers, newsreaders, etc) on whom many researchers rely to produce their data for analysis. »

Les études portant sur des événements vocaux produits en interaction, mais en dehors des tours de parole du locuteur se focalise en grande majorité sur la fonction de *backchannel* (*cf.* Chapitre 2 I.4.1.) que peut avoir ces occurrences. Ce n'est toutefois qu'une de ses nombreuses fonctions. Par exemple Schuller, Eyben, & Rigoll (2008, abstract, p.99) précise que « les vocalisations non-verbales, comme le rire, la respiration, l'hésitation, ou l'accord, jouent un rôle important dans la reconnaissance et la compréhension de la parole conversationnelle humaine et des affects spontanés. »[77]

Une autre fonction majeure de ces occurrences, qu'elles soient produites pendant ou en dehors du tour de parole de l'interactant, est la fonction affective, soulignée par la terminologie « *affect bursts* » (Scherer, 1994). Par ce terme, Scherer (*ibid*, p. 170) réfère à « des expressions très brèves, discrètes et non verbales d'affect, à la fois sur le visage et dans la voix, et provoquées par des événements clairement identifiable »[78].

D'une manière plus générale, nous faisons quant à nous l'hypothèse que ces événements vocaux sont des éléments pertinents pour l'expression des affects, mais également plus largement pour l'expression du *FoT*.

Il est à noter que bien avant de s'intéresser aux *affect bursts*, Scherer s'est d'abord focalisé sur la voix en tant qu'indice affectif, en particulier par la prosodie et la qualité de voix (Scherer & Zei, 1989, p.61). Pour lui, ces « traits [de parole] porteurs d'informations non linguistiques […] apparaissent soit spontanément en caractérisant l'état physiologique concomitant à l'état affectif du sujet parlant, soit de façon intentionnelle en caractérisant les attitudes du locuteur ainsi que ses stratégies d'interaction sociale ».

Ainsi, plus que de simples éléments apparaissant sporadiquement au cours de l'interaction, les événements vocaux peuvent être considérés comme des éléments faisant partie d'un système complexe de nature multimodale, mais aussi comme des éléments dont la nature et les paramètres variés (acoustiques, prosodiques et articulatoires) peuvent être la conséquence physiologique des états affectifs du locuteur lors de leurs occurrences (Scherer, 1994 ; Fredrickson & Levenson, 1998). Gestes articulatoires, états physiologiques et sons produits seraient donc étroitement liés.

[77] Citation originale : « Non-verbal vocalizations such as laugher, breathing, hesitation, and consent play an important role in the recognition and understanding of human conversational speech and spontaneous affect. »

[78] Citation originale : « […] what I call « affect burst », that is, very brief, discrete, nonverbal expressions of affect in both and voice as triggered by clearly identifiable events. »

### I.1.3. Un symbolisme des événements vocaux ?

Approfondissant ce lien entre sons produits et paramètres articulatoires, acoustiques et physiologiques, le courant éthologique du *sound symbolism* (littéralement « symbolisme du son »), est représenté en particulier par Jespersen (1921), Sapir (1925), puis Berlin (1994), Ohala (1994) ou encore Gussenhoven (2002). Ce courant s'intéresse à la relation existant entre forme et fonction. Il montre en particulier qu'il existe un lien motivé entre la forme d'un patron intonatif et sa signification ou sa fonction en réalisant des études comparatives du comportement humain et non-humain (Ohala, 1994). Trois « codes biologiques », les f*requency code, effort code* et *production code,* permettent d'expliquer ce qui est universel concernant l'interprétation de la variation de F0, mais aussi des interprétations plus spécifiques (Gussenhoven, 2002). Il existerait ainsi une certaines icônicité des formes sonores. Un exemple du *frequency code* souvent cité est le fait que la F0 de la voix fournit indirectement une impression de taille à propos de l'individu : une F0 faible donne l'impression d'être gros (« intention » d'attaque, d'une apparence dangereuse), alors qu'une F0 élevée, telle celle des femmes et des enfants, donne l'impression d'être petit (« intention » de paraître sur la défensive).

Selon ce courant (*ibid*, p.48), « les codes biologiques sont ainsi fondés sur les effets de propriétés physiologiques du processus de production sur le signal, mais ces effets ne sont pas automatiques et la physiologie n'est pas suffisante ». Il existe donc un contrôle qui élimine le rapport de cause à effet direct entre physiologie et variation de F0.[79]

### I.1.4. Un classement des expressions sur un axe orienté

D'un point de vue théorique, les événements vocaux et plus globalement les expressions, peuvent également être classés sur un axe orienté. Cette idée est apparue dans la deuxième moitié du XX$^{ème}$ siècle à travers les théories évolutionnistes et a été reprise dans différentes approches, ce qui implique des variations dans les critères de classement et dans sa nature catégorielle ou continue.

Ohala (1996) a établi une taxonomie générale à trois niveaux, issue des théories éthologiques et fondée sur sa distinction entre signe et signal (Chapitre 3 III.). Elle s'applique à tous les types d'expressions d'émotions et d'attitudes, et est donc également pertinente pour classer les événements vocaux. Elle distingue :

[79] Citation originale : « Biological codes are based on the effects of physiological properties of the production process on the signal, but communication by means of the codes does not require that these physiological conditions are actually created. It is not enough to create the effects. That is, the effects are not automatic, but have been brought under control. »

- les événements qui reflètent l'état (psycho-)physiologique du locuteur, qui ne sont informatifs que par inadvertance (*e.g.* la transpiration excessive)

- au niveau fondamental, les « signaux » qui transmettent des messages dont la transmission elle-même a valeur de survie (*e.g.* le sourire). Ces signaux inter-culturels seraient motivés plus par l'effet qu'ils sont susceptibles d'avoir sur l'interlocuteur, que par le propre état (psycho-)physiologique du locuteur (Kraut & Johnston, 1979).

- les signaux proprement dit, par lesquels les humains transmettent leurs attitudes vis à vis de l'interlocuteur, le contenu ou le référent du message, ou encore sur eux-mêmes (incluant l'ironie, l'indifférence, etc.). Ces signaux, probablement appris, peuvent varier énormément d'une culture à l'autre, et même d'un individu à l'autre. Contrairement aux « signaux » des autres niveaux, ils requièrent probablement plus de pragmatique et un contexte linguistique de haut-niveau pour être communiqués de manière appropriée.

Les critères de classement sont donc pour Ohala le type informatif et ainsi l'augmentation de l'intention communicative des items considérés.

Le sociologue Goffman (1981, p.115) « a relevé le fait que certaines *nonword vocalizations* (en quelque sorte "vocalisations non-lexicales"), comme "Uh?" et "Shh!" en anglais, font clairement partie de paroles "directes". Elles sont souvent interchangeables avec un mot proprement dit (*e.g.* dans ces cas là, respectivement "What ?" et "Hush!"). À l'inverse, d'autres, comme "uh" lorsqu'il est utilisé en tant que *filled pause*, appartiennent à un type d'action radicalement différent, à savoir une expression supposée être pure, ce qu'il nomme "*response cry*" et qui correspond à une partie des "expressions naturelles" »[80] (voir Chapitre 5 II.). À partir entre autres de cette observation, il a fait naître le concept d'un continuum entre le langage proprement dit et ce qui est non-linguistique (cité par Wharton, 2003, p.50).

Reprenant cette idée de Goffman, Wharton (*ibid*) a quant à lui établi un continuum dit « showing-saying ». Il illustre sa proposition par le fait qu'un large spectre de comportements permet à un individu de communiquer un ressenti de douleur. Ce spectre va « des expressions faciales complétement naturelles et instinctives, au cri, à l'interjection (anglaise) *ouch*, dépendante de la langue et de la culture, et enfin à l'expression complétement linguistique. Il indique que les interjections ont une part de naturalité et de spontanéité qui suggèrent qu'elles se trouvent *quelque-part entre* le naturel et le linguistique. D'autre part, ces interjections semblent partager avec

[80] Citation originale : « […] note that some [nonword vocalizations] (such as *Uh?* and *Shh!*) are clearly part of directed speech, and often interchangeable with a well-formed word (here *What?* and *Hush!*), but others (such as the *uh* as filled pause) belong to a radically different species of action, namely, putatively pure expression, response crying. »

l'intonation de la voix, les expressions faciales et même les gestes, la propriété d'être en partie codées et en partie naturelles »[81] (*ibid*, p.47).

Afin de concilier ces deux versants des interjections, Wharton proposa (*ibid*, p. 68 - 69) qu'elles se placent à divers endroit d'un continuum des comportements communicatifs, partant de ceux qui fournissent des traces relativement directes d'information - *showing*, à ceux dont tout indice apporté l'est de manière indirecte - *saying*.

Ce continuum des expressions, schématisé Figure 42, est ainsi établi selon plusieurs critères permettant des distinctions entre les deux extrêmes, *showing* et *saying* (pour plus de précisions, voir Wharton, 2003, p.69 à 82) :

- la signification naturelle vs. non-naturelle (au sens de la distinction faite par Grice) ;
- la production délibérée vs. spontanée, avec une transmission d'information intentionnelle, dissimulée ou « accidentelle » ;
- les notions d'icônicité (*cf*. les onomatopées) et de conventionalisation.

Par ailleurs, avec un point de vue diachronique, l'existence de ce continuum est renforcé par le reflet d'une sorte de progression historique qu'il renvoie, laissant apparaître une augmentation progressive de la stylisation / la codification des expressions (p.76), rappelant la notion de « comportements ritualisés » utilisée en éthologie pour décrire l'évolution d'un comportement. Selon Wharton, il existe de bonnes raisons de supposer que certaines interjections proviennent d'expressions naturelles d'émotion, même si la stylisation les sort de la catégorie purement *showing*, ce qui confirme la relation pouvant exister entre émotions et interjections.

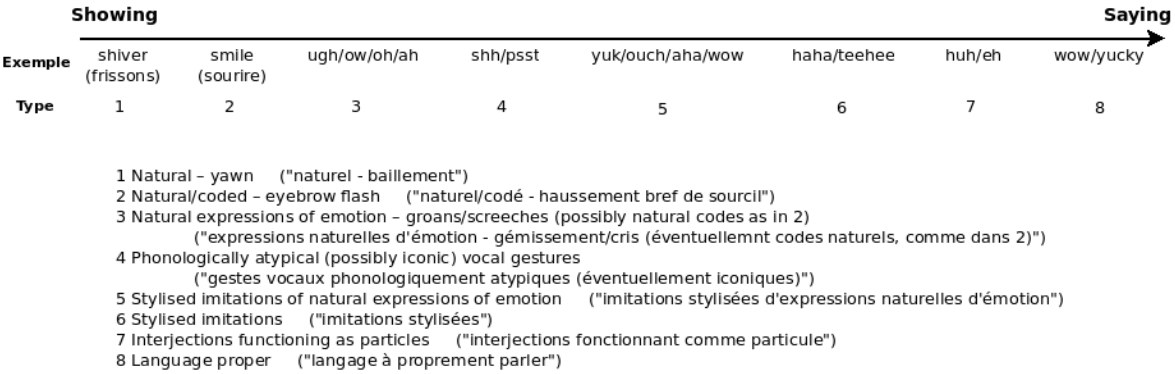



Figure 42: Le continuum "showing-saying" de Wharton (2000, repris en 2003)

[81] Citation originale : « Methods range from allowing someone to see an entirely natural and instinctive contorted facial expression, to a scream such as "aaaargh", to a culture / language-specific *ouch*, to a fully linguistic 'it hurts like hell'. No one would propose that grimaces or screams encode conceptual structure, but communicate they do. Interjections retain an element of naturalness and spontaneity that suggests they fall somewhere *between* the natural and the linguistic. With tone of voice, facial expressions and even gestures, they share the property of being partly coded and partly natural. »

### I.1.5. Vers un axe orienté à valeur phylogénétique

Scherer, se fondant quant à lui sur les théories de Darwin et Trojan, ainsi que sur les travaux sur la communication animal, a mis l'accent sur le caractère phylogénétique de ce continuum des expressions vocales, comme faciales (Scherer, 1994, p.164). Selon lui,[82] « au cours de l'évolution de l'humain, les vocalisations ont été sélectionnées pour servir de signal porteur pour le code le plus important pour la communication humaine : le langage ou plus spécifiquement la parole. […] Ainsi, le mode hautement cognitif, et phylogénétiquement récent, de la communication parlée est greffé sur un système vocal phylogénétiquement ancien, principalement utilisé pour les signaux affectifs et sociaux. Cela a d'importantes implications sémiotiques. Par exemple puisque l'état émotionnel d'une personne affecte en continu sa respiration, sa phonation et son articulation, toute activité de parole fournit des marques continues ou un affichage des états respectifs du locuteur, qui servent d'informations importantes sur le contexte pour l'interlocuteur. » (Scherer, 1994, p. 168)

L'hypothèse que nous proposons rejoint l'axe continu qui relierait, selon Aubergé (2002b) le contrôle involontaire au contrôle volontaire des objets communicatifs (*cf.* Chapitre 2 II.3.2.) : ces événements sonores qui n'entrent pas dans la double-articulation (*i.e.* qui ne procèdent pas à un accès lexical), s'organiseraient sur un axe continu où le contrôle devient volontaire et de plus en plus complexe, jusqu'à la double-articulation. Le vecteur de ce contrôle serait la prosodie : d'abord le contrôle du temps de l'expiration, de sa « force » illocutoire, puis celui de la qualité sonore (pour ce dernier point, *cf.* partie III.2.3. de ce chapitre).

Cet axe classerait alors les items des sons non-phonétiques (« bruits de bouche » variés) aux items pré-lexicalisés (interjections), avec de nombreuses formes intermédiaires (voir Chapitre 7 II.1.2.).

En effet, certains événements vocaux semble avoir un statut intermédiaire dans les langues, comme l'a relevé Vasilescu concernant les hésitations vocaliques :

> « La voyelle d'hésitation n'a pas un statut phonologique. Cependant, elle semble "attirée" par des timbres phonémiques attestés dans le système vocalique d'une langue. » (Vasilescu, Adda-Decker, & Nemoto, 2008, p.211)

[82] Citation originale : « […] in course of human evolution, vocalizations has been selected to serve as the carrier signal for the most important code in human communication -language, or more specifically speech. [...] the phylogenetically recent, highly cognitive mode of speech communication is grafted upon a phylogenetically old vocal system mainly used for affective and social signalling. This has important semiotic implications. For example, because the emotional state of a person continuously affects respiration, phonation, and articulation, all speech activity provides a continuous marking or read-out of the respective state of the speaker, which serves as important context information for the listener. »

Comme l'ont introduit Scherer avec sa considération phylogénétique des signaux affectifs et sociaux, ou Wharton avec la perspective diachronique apportée par son continuum, d'un point de vue évolutionniste, considérer un tel continuum permettrait également d'établir un parallèle avec les théories de l'évolution du langage (*cf.* aussi notamment Deacon, 1997)

Selon la théorie de l'évolution du langage de Jackendoff (1999 ; 2002), dont la partie qui nous intéresse est schématisée dans la Figure 43, l'utilisation de symboles est le facteur le plus fondamental dans l'évolution du langage, une des premières étapes importantes étant l'utilisation volontaire de vocalisations symboliques (ou de d'autres signaux comme les gestes), y compris de manière non spécifique à une situation. Deux étapes en parallèle suivent ensuite : le développement d'un système combinatoire phonologique et la concaténation de symboles, la position de chacun d'eux pouvant apporter du sens (*i.e.* proto-syntaxe).

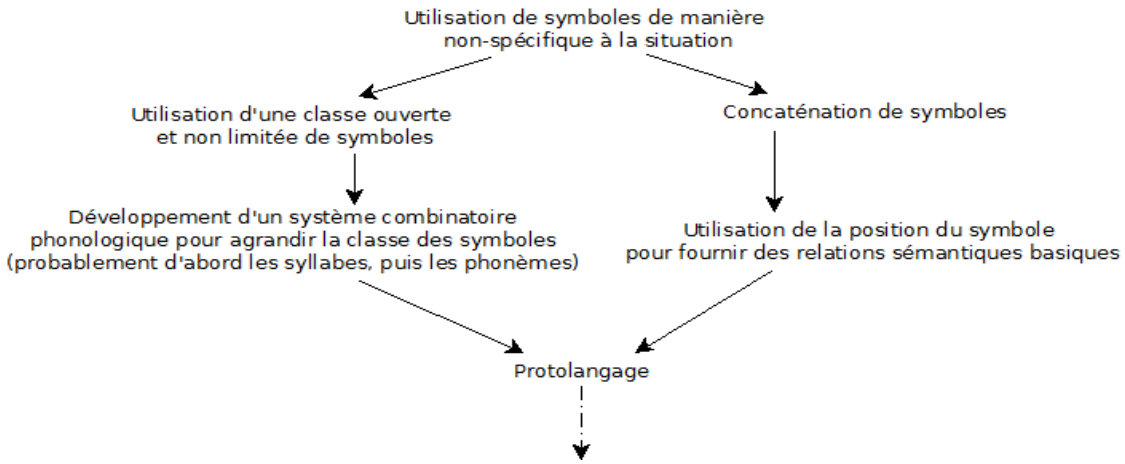



Figure 43: Les premières étapes de l'évolution du langage selon Jackendoff (adaptation et traduction de Jackendoff, 1999, p.273)

Pour Jackendoff (1999), des éléments du langage moderne, certains faisant partie de nos événements vocaux seraient des « fossiles » d'étapes antérieures de l'évolution. Pour exemple, parmi ces énoncés qui n'ont pas de syntaxe, se trouvent d'une part les énoncés associés à un affect intense et soudain (*e.g.* « aïe ! », « wow ! »), exclamations faisant partie des interjections et souvent non communicatif et involontaire ; d'autre part des énoncés spécifiques à la situation, volontaires et avec un but d'agir sur l'autre inhérent, (*e.g.* [ʃ], transcrit « chut » en français).

Avec un point de vue dont l'intérêt est plus porté sur l'expression des affects que sur la linguistique, Simon-Thomas et al. (2009) indiquent que leurs *vocal bursts* ont probablement émergé tôt dans l'évolution du primate et ont été retenus dans le répertoire émotionnel de l'humain. Ils le justifient par une observation comparatiste : les jugements portant sur les *vocal bursts* rejoignent ceux des catégories de

vocalisations émotionnelles décrites chez les primates (appréciation d'aliments, roucoulement, sourire -*i.e.* catégorie des vocalisations pro-sociales-). Cela conforte la prémisse que les vocalisations émotionnelles des humains ont évolué vers des buts communicatifs comparables (Snowdon, 2003), et clarifie la qualité et la résolution de la capacité de la voix humaine, à communiquer de l'information émotionnelle en dehors de la parole.[83]

## *I.2. Apports réciproques entre études théoriques et technologies*

Étudier les caractéristiques vocales impliquées dans les processus de reconnaissance des émotions nécessite (Grandjean & Baenzinger, 2009, p. 129-130) : soit d'établir des corrélations multiples entre les caractéristiques acoustiques de l'expression, et l'émotion qui lui est attribuée perceptivement ; soit d'étudier la perception émotionnelle d'expressions non naturelles. Dans ce dernier cas, les expressions non-naturelles sont, la plupart du temps, des signaux modifiés en éliminant ou masquant une partie de leur information (en particulier lexicale), par exemple par filtrage, ou alors des expressions (re)synthétisés en manipulant certaines de leurs caractéristiques acoustiques. Pourtant Les événements vocaux, par leurs natures même, peuvent être des éléments portant des paramètres prosodiques spontanés, purs ou combinés, sans contenu lexical (Kompe, 1997). Ils sont ainsi une réponse naturelle à cette contrainte méthodologique inhérente à l'étude des expressions vocales des émotions.

Les événements vocaux permettent donc d'isoler de la prosodie spontanée à fonction émotionnelle (mais probablement pas uniquement) sur des éléments ponctuels et brefs (jusqu'à quelques secondes au maximum) (Scherer, 1994).

De plus, malgré les incidences de l'état physiologique du locuteur sur sa production de sons, la signification de l'intonation réside également dans le contrôle qu'a le locuteur sur les formes produites. Ainsi, connaissant tous deux les relations existant entre formes et fonctions/significations, le locuteur et l'auditeur produisent les formes de manière appropriée à la situation :

> « Un aspect important de cette conception de la signification de l'intonation est qu'alors que la nature des significations est liée à la manière que nos organes vocaux adopte pour produire la variation de F0, il n'y a pas d'implication quant au fait que les conditions physiques à la base de ces significations aient besoin d'être présentes pour

[83] Reformulé à partir de la citation originale : « Many of the vocal bursts documented here are likely to have emerged early in primate evolution and been retained in the human emotional repertoire. Judgment data from these vocal bursts dovetail with emotion vocalization categories described in primates, such as specific calls that indicate the presence of food (savoring) or affiliative-caregiving coos and grins (pro-social), supporting the premise that humans have evolved vocalizations of emotion for comparable communicative purposes (Snowdon, 2003). These patterns of judgment begin to clarify the quality and resolution of the human voice's capacity to communicate emotional information outside of speech. ». (Simon-Thomas et al, 2009, p. 844)

> créer les formes. Locuteurs et interlocuteurs savent quelles sont ces relations entre formes et significations, et produisent les formes de la manière qui convient. »[84] (Gussenhoven, 2002, p.51

Cela n'est pas sans rappeler la théorie de la pertinence de Sperber et Wilson. Cette dernière est fondée sur le principe que les humains sont faits de telle manière à chercher de l'information pertinente, la pertinence de l'information étant définie en termes d'effets cognitifs gagnés et d'efforts produits pour traiter l'information : plus les effets cognitifs gagnés sont grands et les efforts de traitement faibles pour acquérir ces effets, meilleure est la pertinence de l'information (Sperber & Wilson, 1989). Ainsi, en choisissant, volontairement ou non, de produire les formes (notamment d'événements vocaux dans notre cas) les plus appropriées à la situation de la communication, les différentes parties de l'interaction respectent les principes de la théorie de la pertinence. Les événements vocaux produits spontanément par un humain, même en interaction personne-machine, sont donc, selon ces principes, toujours appropriés au contexte, à la situation.

Ainsi, en IHM, il serait plus facile et plus pertinent de générer de manière correcte ces événements vocaux brefs (*e.g.* des « euh » au bon endroit et avec une prosodie correcte), qu'une longue réponse (*e.g.* un énoncé signifiant approximativement « j'en doute, mais bon je vous le dis quand même ») adaptée à la situation par tous ses paramètres (lexicaux, syntaxique, phonétique, mais aussi d'organisation temporelle, de prosodie, etc.). C'est ce que Ward a souligné concernant les items conversationnels :

> « Plus généralement, la communication utilisant des *grunts* conversationnels est certainement préférable à des phrases entières, puisque les *grunts* sont un moyen concis et informel d'indiquer les aspects « meta- » et attitudinaux de l'interaction, qui sont des parties importantes de l'interaction, même si elle est formelle. »[85] (Ward, 2000b, p.572)

L'utilisation de ces événements est d'autant plus pertinente, que l'humain ne peut généralement pas parler et écouter en même temps, alors qu'il est capable de percevoir et interpréter les *grunts* en même temps qu'il parle. Les *grunts* forment donc un canal séparé dont l'utilisation n'interfère pas avec le canal principal.

[84] Citation originale : « An important aspect of the present conception of intonational meaning is that while the nature of the meanings is related to the way our speech organs produce pitch variation, there is no implication that the physical conditions that lie at the basis of these meanings need to be present in order to create the forms. Speakers and listeners know what these form-function relations are, and will produce the forms in the way they see fit. »

[85] Citation originale : « More generally, communication using conversational grunts may be preferable to full sentences as a concise and informal way to handle the attitudinal and 'meta' aspects of interaction, which are important parts in all but the most formalized interactions. »

Les événements vocaux jouent un rôle indéniable en interaction. Les modéliser permettrait d'améliorer les systèmes de reconnaissance de la parole et les autres technologies de l'*Affective Computing* (Pelachaud et al., 2005), par exemple en permettant des « interactions concises, régulières et détendues »[86] (Ward, 2000b, abstract).

Concernant le domaine plus restreint des systèmes de dialogue, Ward, Rivera, Ward, & Novick (2005) ont effectué une comparaison entre interaction humain-humain et interaction personne-machine. Pour eux, et pour ceux qui construisent des systèmes de dialogue, l'intérêt de chercher à imiter les performances d'une interaction entre humains n'est pas un but en soi. Pourtant, les systèmes de dialogue actuels impliquent des formes d'interactions qui ne semblent aujourd'hui ni familières, ni naturelles (Heisterkamp, 2003). D'après l'étude de Ward, Rivera, Ward, & Novick (2005), un des principaux points à améliorer, est que le système ne fournit pas de *feedbacks* variés et complètement appropriés (entre autres au niveau de l'organisation temporelle et de sa prosodie) à la situation, y compris à l'état dans lequel se trouve l'interlocuteur (*e.g.* a-t-il besoin d'être rassurée ou non?). C'est pourquoi les *feedbacks* sont particulièrement étudiés ces dernières années (Pelachaud et al., 2005 ; Schröder et al., 2008 ; Heylen, 2007).

De surcroît, les *feedbacks* pourraient jouer un rôle bénéfique pour l'utilisateur d'une technologie vocale, en augmentant son efficacité lors de certaines tâches, par des félicitations, des encouragements, etc. (Ward, 2000b).

Produire des *fillers* et *disfluency markers* permettrait également de réduire facilement la vitesse de transmission des informations, et ainsi de faciliter leur compréhension. Il s'agit de l'objet d'étude d'Adell, Bonafonte, & Escudero-Mancebo (2010). Cela serait utile pour attirer l'attention de l'utilisateur sur certaines informations en particulier.

Au niveau de la reconnaissance de la parole, et comme l'a souligné Adda-Decker (2006), citée par Vasilescu, Adda-Decker, & Nemoto (2008, p.201) :

> « Dans le cadre de la transcription automatique de la parole, la prise en compte d'un flux acoustique continu implique la modélisation de la totalité des phénomènes présents dans le signal de parole : des mots associés à une représentation phonologique dans le dictionnaire de prononciations, mais également des respirations, des hésitations, des fragments de mots peu ou mal articulés, etc. »

Dans un premier temps, détecter les *grunts* sans même avoir à les interpréter, permettrait de pouvoir les ignorer en tant que parole, d'autant plus qu'ils sont parfois gênant lorsqu'ils interfèrent avec la reconnaissance de mots proches, ou, dans certaines applications téléphoniques, lorsqu'ils couvrent la parole de l'autre partie,

[86] Citation originale : « […] to achieve concise, smooth and relaxed interactions »

voire même lorsqu'ils sont interprétés sans raison (comme commande vocale dans les applications commerciales actuelles).

Par la suite, comprendre la signification des *grunts* conversationnels en système de dialogue, en tant que *backchannel,* pourrait permettre aux utilisateurs de gérer la conversation avec des protocoles humains (*e.g.* la gestion naturelle des tours de parole) (Ward, 2000b).

En 2004, Ward a relevé des éléments comportant un paramètre qu'il a noté « loudness », qui caractérise les événements peu audible. Selon lui, ces éléments particuliers n'ont certainement pas de « fonction pragmatique », mais ils pourraient être utiles pour l'étude en temps réel des processus cognitifs tels qu'ils apparaissent en lien avec le dialogue, c'est-à-dire une partie de notre *FoT*.

D'autre part, Vasilescu, Adda-Decker, & Nemoto (2008) ont étudié les caractéristiques acoustiques et prosodiques des hésitations vocaliques de l'anglais américain, de l'espagnol et du français. Ils ont montré que ces éléments des trois langues partagent certains traits tels que la durée et la hauteur. Des spécificités liées à la langue sont par contre relevées, notamment en terme de timbre vocalique. Ainsi ce type d'événements (ainsi que d'autres événements vocaux de notre corpus peut-être) pourraient être des éléments pertinents pour l'identification automatique des langues.

Enfin, identifier et interpréter les événements vocaux des bases de données de parole apporterait un gain d'information, les utilisateurs ayant parfois besoin de chercher de l'information fournie par les *grunts*, comme le scepticisme, l'amusement, la décision ou encore la présence d'une information importante.

Les événements vocaux pourraient donc jouer un rôle dans la perception du *FoT.* De même, ils pourraient être les témoins d'une évolution de la gestualité vers le langage par leur gradation vers ce dernier. L'expression du *FoT* semble en conséquence ancrée dans une multimodalité complexe.

D'autre part, ces événements augmenteraient la naturalité et la compréhension, à la fois pour les technologies d'analyse, de synthèse, de transcription de données, et pour les systèmes de dialogue. Il serait donc utile, en particulier pour les technologies de parole et les agents virtuels) d'être capable de reconnaître automatiquement et de générer ces événements dits non langagiers. Cependant, leur(s) nature(s) acoustique(s) et prosodique(s), autant que leurs fonctions, leur leur organisation temporelle, et leur ancrage dans le langage sont encore peu décrits. Pourtant, la pertinence du comportement communicatif passe certainement par les stratégies de génération et d'identification des événements qui « encadrent » la communication langagière.

# II. Aperçu des études sur les événements vocaux et de leur problématique

Ce à quoi nous nous référons en parlant des événements vocaux correspond à un ensemble d'éléments de nature très variée. Leurs points communs sont leur modalité perceptive au moins acoustique, et leur apparition en interaction face à face. Dans la littérature, notre objet d'étude concerne donc les interjections au sens commun du terme[87], étendues à d'autres éléments tels que des vocalisations diverses (gémissement, rire, cris, etc.) et autres *non-lexical speech sounds* (littéralement « sons de parole non-lexicaux » - Ward, 2000b).

## *II.1.De l'étude des interjections...*

Les premiers essais parlant des origines et des fonctions des interjections, sont des écrits de psychologie du langage dont l'intérêt portait alors essentiellement sur la question de l'origine du langage humain. Par exemple Kleinpaul (1888) parlait de *nature and feeling sounds* (littéralement « sons de la nature et des sentiments »). Il cherchait entre autres à vérifier que ces sons étaient produits de manière identique par des locuteurs de différentes cultures (cité par Scherer, 1994, p.171).

D'autre part, il a distingué les interjections ou exclamations produites spontanément et exprimant un état émotionnel, des appels et cris émis intentionnellement pour des raisons communicatives. Il a cependant précisé qu'« un même son (*e.g.* [o]) peut être utilisé à la fois comme interjection, et comme signal vocal délibéré »[88] (Kleinpaul, 1888, cité par Scherer, 1994, p.171).

Par la suite, Wundt (1900) fut l'un des premiers à s'intéresser aux interjections dans leur lien avec le langage. Il distingua deux types d'interjections :

- les interjections « primaires », sons de la nature, réminiscences d'une période prélinguistique et dont la seule fonction est d'interrompre la continuité de la parole ;
- les interjections « secondaires », assimilées à la langue.

Selon Wundt, ces dernières remplaceraient les interjections primaires dans le développement linguistique et culturel. Ainsi, leur nombre dans une langue augmenterait avec le niveau de civilisation de la culture de ses locuteurs, et des contraintes imposées par les normes et les mœurs sociales sur l'expression verbale.

[87] Définition du TLFI : « Mot invariable, autonome, inséré dans le discours pour exprimer, d'une manière vive, une émotion, un sentiment, une sensation, un ordre, un appel, pour décrire un bruit, un cri ».

[88] Citation originale : « the same sounds, for example "o" can be used not only as an interjection, but also as a deliberate vocal signal. »

Par exemple dans une culture où un fort contrôle est exercé sur l'expression émotionnelle, un individu développerait une interjection secondaire comme « aïe » pour remplacer une interjection primitive de type « aghghgh » (Wundt, cité par Scherer, 1994, p.171-172).

Une cinquantaine d'année plus tard, Kainz (1962) reprit cette distinction entre interjections primaires et secondaires, mais la contrasta également avec les « sons de la nature », « cris et soupirs informes, qu'il considère comme de pures vocalisations réflexives, servant seulement à communiquer des émotions »[89] (cité par Scherer, 1994, p.172). Les éléments de cette dernière catégorie, produits non intentionnellement, et ne pouvant être décrits de manière stable avec des symboles phonétiques, ne peuvent donc, selon Kainz, faire partie des interjections de la langue (les interjections primaires et secondaires). La raison est la nécessité pour cela d'« une stylisation et d'une conventionalisation [de la forme sonore], qui a modelé ces sons de manière à ce qu'ils correspondent aux règles phonologiques d'une langue donnée »[90] (*ibid*, p.172).

Nous choisirons quant à nous de nommer interjections les items référencés par la plupart des dictionnaires grand public. Ce sont les items à qui l'usage a donné une transcription orthographique, c'est-à-dire une adéquation au système phonologique et une fonction consensuelle. Les autres items, même phonétiquement inscrits dans la phonologie de la langue(ici le français) seront nommés « bruits de bouche » (*cf.* partie III.2.).

Par la suite, deux approches théoriques sur les interjections se sont souvent opposées, le principal enjeu du débat étant de déterminer si les interjections forment une classe lexicale, et si elles font partie de la langue. Pour des détails sur les enjeux du débat et les arguments de chaque position, se reporter par exemple à Wharton (2003), Li (2005), ou Cruz (2009).

D'un côté se trouve le paradigme sociolinguistique, représenté en particulier par Goffman (1981a), qui continua à développer l'idée d'évolution, mais sans l'aspect linguistique. Pour Goffman, une interjection (comme « aïe », « oops », « wow », « ouh la ») est un acte ritualisé au sens éthologique du terme. Selon cette perspective, les interjections, qu'il nomme aussi *response cries* ne feraient donc pas partie du langage : ce serait des « non-mots ». Il les analyse ainsi selon le rôle socio-communicatif qu'elles jouent, sans aucun contenu linguistique (décrit par Wharton, 2003).

[89] Citation originale : « amorphous screams and sighs, which he considered to be sheer reflexive vocalizations serving only to release emotions. »

[90] Ciatation originale :

De l'autre côté, il existe une deuxième approche théorique des interjections, appelée « théorie sémantique » par Li (2005), ou *conceptualist view* par Wharton (2003). Elle considère quant à elle les interjections comme une des parties du discours, et les analyse sémantiquement de manière complexe. Les interjections sont, selon cette approche, « sémantiquement riches, et ont une structure conceptuelle précise qui peut être expliquée »[91] (Wilkins, 1992, p.120). S'inscrivant dans cette approche, Ameka (1992) reprend la distinction entre interjections « primaires » *vs.* « secondaires » introduite par Wundt, mais pas tout à fait avec la même distinction :

- les interjections primaires ne sont pas utilisées d'une autre manière, ce sont des petits mots ou « non-mots » qui peuvent constituer un énoncé par eux-mêmes. Ce type d'interjection a tendance à être phonologiquement et morphologiquement « anormal », c'est-à-dire fait de sons et de séquences sonores qui ne peuvent être trouvées dans les autres parties du langage (*e.g.* les interjections anglaises *pm!* et *sh!* qui ne contiennent aucune voyelle) ;
- les interjections secondaires appartiennent à d'autres classes de mots en fonction de leur sémantique. Ce sont des interjections uniquement car elles peuvent se rencontrer par elles-mêmes et de manière non-elliptique, en tant qu'énoncé à un seul mot. De plus, elles réfèrent dans cet usage à des actes mentaux.

Ameka (1992) a par ailleurs établi différents critères permettant de classer un item dans les interjections :

- du point de vue structural / morphologique, les interjections ne prennent habituellement pas de flexions ni de dérivations dans les langues ;
- au niveau sémantique, ce sont des gestes vocaux relativement conventionnels qui expriment l'état mental, l'action, l'attitude du locuteur, ou encore sa réaction à une situation ;
- du point de vue pragmatique, il s'agit d'items liés au contexte, qui encodent les attitudes et intentions communicatives du locuteur.

Pour Wierzbicka (1992), ce sont les deux premiers critères donnés par Ameka qui déterminent la classe grammatical des interjections.

L'approche sémantique de l'interjection s'est assez répandue dans les domaines de la linguistique. L'approche linguistique a souvent défini l'interjection de manière large :

> « comme un élément isolé du discours, invariable et indépendant des autres propositions et constituants […]. Sur le plan sémantique elle traduit la charge émotionnelle du locuteur en permettant l'expression d'une large gamme d'états d'âme et de réactions affectives. » (Contini, 1989, p.320)

[91] Citation originale : « semantically rich and have a definite conceptual structure which can be explicated ».

Son intérêt est de permettre, de par l'indépendance comme critère définitionnel de l'interjection, de s'affranchir du débat cité précédemment. L'objectif est alors de réaliser un inventaire formel et sémantique des interjections, voire prosodique comme par exemple dans l'étude de Contini (*ibid)* sur les interjections en Sarde.

Plus récemment, une approche alternative, cognitive-pragmatique, s'est développée, dans laquelle l'interjection est considérée à mi-chemin entre les « sons naturels » et les éléments linguistiques. Elle reproche entre autres à l'approche sémantique le fait de vouloir décomposer sémantiquement la signification des items, et de ne pas prendre en compte le contexte, qui pourtant peut changer le sens d'une même interjection. Elle est fondée sur la théorie de la pertinence de Sperber & Wilson (1989 ; *cf.* aussi I.2.), et sur des concepts complexes et subtiles de linguistique et de philosophie du langage, tels que les actes de langage (Austin, 1975) ou les conditions de vérité d'un énoncé (Grice, 1989). Pour approfondir les subtilités de cette approche et les désaccords existants au sein de ses partisans, se reporter à Wharton (2003), Li (2005) ou Cruz (2009).

Sans être toujours explicitement utilisée, cette approche trouve résonance dans certains travaux actuels, plus applicatifs. Par exemple en *Affective Computing*, Poggi (2008), a adapté la définition d'interjection de l'approche sémantique à son champ d'étude, en lui ajoutant la notion de *codified signal* (littéralement « signal codifié »), c'est-à-dire « un signal perceptible […] qui est lié de manière stable, dans les esprits des locuteurs d'une langue, à la signification d'un acte de langage »[92] (*ibid,* p.171).

### *II.2…. à celle des événements vocaux*

Dès le début des études portant sur les interjections, une intuition quant à leur gradation d'un point de vue linguistique et intentionnalité apparut (*cf.* les différentes distinctions entre interjections primaires *vs.* secondaires). Il en est de même de l'existence d'éléments à la limite de l'interjection, mais sans aucun statut linguistique (*cf.* les « sons naturels » de Kainz -voir la partie précédente-).

Étant donné notre méthodologie empirique (*cf.* Chapitre 3), nous nous attachons d'abord à décrire tout ce qui est perceptible, en l'occurrence, dans cette partie de nos travaux, tous les événements vocaux, sans nous focaliser particulièrement sur les « pures » interjections.

Depuis les dernières décennies du XX$^{ème}$ siècle, les éléments non-lexicaux isolés ainsi que leurs formes phonétiques et leur prosodie sont étudiés soit pour leurs fonctions émotionnelles en phonostylistique (Fónagy, 1991 ; Léon, 2005), soit pour leurs

[92] Citation originale : « a perceivable signal […] which is linked in a stable way, in the minds of the speakers of a language, to the meaning of a speech act. »

fonctions pragmatiques dans le dialogue. Ils deviennent alors dans ce dernier cas une partie du discours, une catégorie lexicale désignant généralement les interjections (Ameka, 1992 ; Wierzbicka, 1992 ; Ward, 2000b ; Wharton, 2003 ; Campbell, 2004).

Dans les approches récentes, et comme dans les recherches sur les émotions (*cf.* Chapitre 1), la dichotomie entre émotion et cognition disparaît. Les études concernant les *affects bursts* (Scherer, 1994) ou les *mind markers* (Poggi, Pelachaud, & Magno Caldognetto, 2004) en sont l'illustration. Ainsi, les critères qui servent actuellement à catégoriser les événements vocaux et même à déterminer un objet étude, sont variables.

### II.2.1. Terminologies utilisées et catégorisation des événements vocaux

L'ensemble de ces événements vocaux a fait l'objet d'une littérature plus dispersée que celle portant sur les interjections. Cet ensemble a ainsi été décrit comme une classe d'items hétérogène, et en utilisant des terminologies variées, dépendantes du domaine de l'auteur.

Dans une terminologie évolutionniste, ce qui est appelé *grunts* (Ohala, 1996 ; Ward, 2000a ; Campbell, 2004) se rapporte à des grognements pré-verbaux, dans la lignée terminologique des *response cries* (Goffman, 1981b) et des *vocal segregates* (Trager, 1958).

À différents niveaux plus fonctionnels, et en tenant plus ou moins compte de la forme des objets considérés, les auteurs parlent de *bursts* (Simon-Thomas, Keltner, Sauter, Sinicropi-Yao, & Abramson, 2009) ou *affect bursts* (Scherer, 1994, Schröder, 2003), *filled pauses* (Maclay & Osgood, 1959), *fillers* (Sadanobu, 2004 ; Candea, Vasilescu, & Adda-Decker, 2005 ; Dijkstra, Krahmer, & Swerts, 2006), d'hésitations vocaliques (Vasilescu, Adda-Decker, & Nemoto, 2008), d'interjections dans un sens plus ou moins large (Ameka, 1992 ; Wharton, 2003 ; Poggi, 2008 – *cf.* partie précédente-), ou encore de *nonverbal emotional vocalizations* (Sauter, Eisner, Ekman, & Scott, 2010).

Ces terminologies se rapportent plus ou moins aux mêmes objets. Ainsi, le champ des événements vocaux forme un objet d'étude tellement large que la plupart du temps, les chercheurs utilisent des critères pour le restreindre. Pour cela, soit ils se focalisent sur les événements liés à une ou plusieurs fonctions particulières, soit la restriction est dictée par la nature des objets. Ces paramètres deviennent alors les critères qui déterminent *a priori* les items considérés.

La nature des objets est souvent un critère dans les approches phonétiques, phonologiques et acoustiques. Elle est en particulier utilisée pour déterminer les

interjections ou pour des éléments particuliers comme les rires (*cf.* l'étude de Loyau (2007), ou la base de données AVLaughterCycle de Urbain et al. (2010).

Quant aux approches dont le critère est fonctionnel, elles fixent une fonction à laquelle sont éventuellement associés quelques critères de forme (*e.g.* uniquement les sons vocaliques) et/ou de position (*e.g.* occurrences pendant le tour de parole ou non), qui restreignent plus ou moins l'objet. Ensuite, elles en relèvent les événements correspondants, parfois automatiquement, comme par exemple Vasilescu, Adda-Decker, & Nemoto (2008) pour les fillers, ou Sajjanhar & Ward (2006) pour les signaux de *backchannel*. Ces approches fonctionnelles sont souvent utilisées :

- dans les approches pragmatiques, de par leur fondement théorique de recherche des fonctions du langage en situation de dialogue ;
- dans le domaine de la modélisation et des ACAs, car l'utilisabilité des résultats est immédiate, en construisant des « lexiques » mettant en correspondance geste et/ou événement vocal de nature particulière, et fonction (lié à la communication).

Les fonctions étudiées sont par exemple l'hésitation vocalique, la fonction de *backchannel*, et plus généralement de *feedback*, l'expression émotionnelle/affective, les *filled pauses* et autres « *disfluency markers* » (Brennan & Williams, 1995), ou encore les *mind markers* (Poggi, 2002). Remarquons que toutes ces fonctions ont leurs événements repérables à partir de l'étude de l'état du locuteur, ou de l'analyse du contexte conversationnel (en particulier pour les *filled pauses*).

Dès que la description ou la modélisation cherche à devenir plus globale, un problème apparaît avec ce type d'approche : les formes sont multi-fonctionnelles. Même au sein de sa catégorie des « vocalisations affectives » (*affect vocalizations*), c'est-à-dire des événements qui, par définition, sont à fonction affective, Scherer (1994, p.172-173), relève que les formes semblent presque toutes être utiles aux fonctions sémantiques, syntaxique, pragmatique et dialogique du comportement non-verbal en conversation.

Une autre problématique découlant de l'utilisation d'un de ces critères est que cela influence les transcriptions (préciser avec Ward: issue in transcription...). L'étude idéal consisterait à partir des données, recenser les événements sans tri *a priori*, et enfin effectuer des analyses. Pourtant ce type d'étude est rare.

### II.2.2. Les technologies et leur approche souvent plus globale et/ou plus formelle des événements vocaux

Les enjeux des technologies du langage naturel et de la parole, du domaine de l'*Affective Computing,* et en particulier celui des ACAs, nécessitent d'observer des situations écologiques et réalistes (voir Chapitre 2 ). Or dans ces situations, les

événements vocaux rencontrés peuvent être de toutes sortes, et toutes les fonctions sont représentées :

> « Un tel système, cependant, ne fonctionnera pas bien pour la parole conversationnelle spontanée […]. Ceci est du à la variété des sons non-verbaux et aux irrégularités rencontrées dans la parole spontanée. Cela inclue les *disfluencies* (pauses remplies ou non, corrections et mots incomplets), les interjections (*e.g.* rire, pleur, accord / désaccord : "aha / ah ah"), les bruits humains (*e.g.* bâillement, raclement de gorge, respiration, occlusion et click (*smacking*) toux, reniflement) et les autres sons tel que l'arrière plan de la conversation ou le bruit ambiant (Ward, 1991). »[93] (Schuller, Eyben, & Rigoll, 2008, p.99-100)

C'est pourquoi l'approche utilisée dans les domaines technologiques est souvent plus globale. Dans les systèmes de reconnaissance, même si l'objet d'étude est restreint, sa description formelle et/ ou situationnelle doit être suffisante pour le distinguer du flux de parole et des autres événements audibles. L'approche doit prendre en considération ces derniers, et reste donc globale. Les objets les plus étudiés en vue de reconnaissance automatique sont les *filled pauses* (*e.g.* Goto, Itou, & Hayamizu (1999) sur des paramètres acoustiques, ou Vasilescu, Adda-Decker, & Nemoto (2008)), et les rires (*e.g.* Campbell, Kashioka, & Ohara (2005), avec des méthodes statistiques HMM).

Pendant longtemps, le but de ces études a été de détecter ce type d'éléments afin de ne pas les prendre en compte dans l'analyse, et ainsi d'améliorer la robustesse de la reconnaissance de parole (Schultz & Rogina, 1995). Une prise de conscience de l'information portée par ces éléments a ensuite eu lieu, et l'intérêt s'est alors porté sur leur reconnaissance et leur identification (*e.g.* Lickley, Shillcock, & Bard (1991) ou Campbell (2007)). Ainsi, depuis quelques années, les systèmes s'étendent à des événements vocaux plus variés, étudiés en tant qu'apport d'information (*e.g.* Schuller Eyben et Rigoll, 2008).

Étant donnée la variété des formes à étudier, les systèmes de reconnaissance prennent de plus en plus en compte des paramètres non plus formels, mais temporels et situationnels. Par exemple Sajjanhar & Ward (2006) ont développé un système d'étiquetage automatique de *backchannel* utilisant uniquement les patrons de parole et silence des interlocuteurs et le contexte local des éléments.

[93] Citation originale : « Such a system will, however, not work well for spontaneous, conversational, speech […]. This is due to various non-verbal sounds and irregularities encountered in spontaneous speech. These include disfluencies (filled and unfilled pauses, corrections and incomplete words), interjections (*e.g.* laughing, crying, agreement/disagreement : "aha/ah ah"), human noises (*e.g.* yawning, throat clearing, breathing, smacking, coughing, sneezing) and other sounds like background conversation or noise (Ward 1991). »

En parallèle, dans un objectif de synthèse (vocale ou multimodale), il est indispensable de ne plus « trier » les événements à étudier sur des critères fonctionnels, afin de modéliser un comportement complet et cohérent.

### II.2.3. Vers une approche empirique des événements vocaux

Malgré l'évolution des approches vers une globalité, liée en particulier aux nouveaux objectifs technologiques, peu d'études utilisent aujourd'hui une méthodologie empirique, c'est-à-dire qui part des objets sans les trier, puis les analyse.

Pourtant, ce type de méthodologie, proche de l'éthologie (*cf.* Chapitre 3) serait, comme pour les modalités visibles, particulièrement bien adapté aux enjeux actuels de l'étude des événements vocaux. Ainsi, pour Ohala (1996, p.1812), l'éthologie est une « ressource précieuse pour les données, méthodes et théories liées à la question de comment les attitudes et les émotions peuvent être transmises dans la parole et la kinésique qui l'accompagne »[94]. De plus pour lui (*ibid*, p.1812) :

> « Les descriptions éthologiques ont le plus de chance d'apporter une théorie unifiée de ces comportements. »[95]

L'approche éthologique permet ainsi à Ohala de vérifier ses hypothèses sur les trois codes biologiques (*cf.* partie I.1.3. de ce chapitre).

En dehors des études éthologiques, seuls quelques chercheurs ont adopté des approches empiriques pour étudier les événements vocaux (*e.g.* Scherer (1994), Sauter, Eisner, Ekman, & Scott (2010), Simon-Thomas, Keltner, Sauter, Sinicropi-Yao, & Abramson (2009) ; Ward (2006)). Parmi eux, (Ward, 2004) a mené des recherches qui ont permis de recenser les différentes fonctions de ces événements après les avoir classés. Il s'est également intéressé, en parallèle (Ward, 2000a), aux différentes méthodes d'annotation pouvant être adoptées.

Notons que les études dont les approches sont les plus globales et / ou empiriques, donnent souvent aux événements vocaux, une terminologie / définition par exclusion : ce sont des « événements (de parole) », des « sons » ou des « vocalisations » non-verbaux et/ou non-lexicaux (*e.g. non verbal speech events* et *non-word sounds* pour Campbell (2007), *non-verbal vocalizations / sounds* pour Schuller, Eyben, & Rigoll (2008), ou *non-lexical speech sounds* pour Ward (2000b).

[94] Citation originale : « valuable ressource for data, methods and theories relevant to the question of how attitudes and emotions might be conveyed in speech and its accompanying kinesics. »

[95] Citation originale : « Ethological accounts have the best chance of providing a unified theory of these behaviors. »

Quant à l'approche que nous adoptons, elle se veut empirique et ne sélectionne pas les objets étudiés selon un critère fonctionnel. Elle s'intéresse ainsi à l'ensemble des événements vocaux avec pour seul critère qu'ils soient perceptibles.

## *II.3. Problématique des événements vocaux*

Dans une perspective de communication, située sur le sujet, et dans le temps et le contexte de la situation, l'étude de ces événements vocaux ne peut trouver d'intérêt que si la pertinence communicative de ces événements est au minimum démontrée, et à terme mesurée et qualifiée.

Nous relaterons Chapitre 7 une expérience préliminaire sur les données (Signorello, Aubergé, Vanpé, Granjon, & Audibert, 2010), posant les jalons d'une démonstration de la potentielle pertinence langagière et culturelle des événements vocaux. Nous n'avons pas encore pu, plus précisément, identifier quelles valeurs du *FoT* sont désignées par ces événements.

Toutefois, un point central qui oriente notre travail est la nature du contrôle par lequel sont produits ces gestes vocaux : l'hypothèse de la prise de contrôle volontaire *vs.* contrôle involontaire (Aubergé, 2002a).

Il s'agit ainsi de séparer les signaux qui :

1) ne révèlent rien (bruits non pertinents informativement) ;
2) ne révèlent rien directement mais dont la répartition dans l'espace communicatif (les motifs temporels) sont révélatrices d'indices d'états physiologiques, émotionnels, attitudinaux, mentaux, etc., à travers leur trace comportementale ;
3) sont des indices « récupérés » d'états émotionnels ou physiologiques, mais ne sont pas des signaux puisque ce sont des conséquences d'états audibles ;
4) sont des signaux automatiquement déclenchées (*e.g.* d'états émotionnels) ;
5) sont des signaux déclenchées volontairement, et par conséquent intentionnellement.

Les descriptions que nous mènerons plus loin se situent dans une approche impressionniste, ou du moins se fondent sur une écoute très attentive des signaux, sans référence à leur situation, et en minimisant les risques d'interprétation qui fausseraient l'écoute.

# III. Notre objet d'étude et sa description : une approche empirique des événements vocaux

Pour six sujets de notre corpus, nous avons dans un premier temps étiqueté un grand nombre d'événements vocaux, pouvant être très différents de part leurs caractéristiques acoustiques / articulatoires. Comme pour les événements gestuels / faciaux, nous avons cherché à savoir quels paramètres pouvaient être communicativement pertinents pour l'expression du *FoT*. Cette pertinence peut être de l'ordre des caractéristiques intrinsèques de l'événement vocal considéré, comme en lien avec la situation, le contexte de l'interaction, c'est-à-dire de l'ordre de la pragmatique ou de la psychologie comportementale. C'est pourquoi dans un premier temps, nous n'avons négligé aucun paramètre et avons donc étiqueté un grand nombre d'éléments, de manière à laisser une marge de manœuvre pour les analyses futures.

Après une revue de quelques considérations théoriques et pratiques concernant l'étiquetage d'éléments pragmatiques et/ou en lien avec la situation / le contexte, ainsi que le choix des sujets, nous détaillerons les critères utilisés lors de l'étiquetage intrinsèque des événements vocaux, et nous terminerons par les difficultés rencontrées, en lien avec notre méthodologie empirique, et les interrogations soulevées par l'étiquetage.

## *III.1. Quelques considérations théoriques et pratiques*

### III.1.1. Étiquetage d'éléments d'ordre temporel

Nous considérons ici des éléments d'ordre temporel liés au temps communicatif (c'est-à-dire aux tours de parole du sujet), l'étiquetage des éléments du temps événementiel ayant été traité Chapitre 3 IV.3..

L'étiquetage de la tâche dans laquelle se trouve le sujet (*cf.* Chapitre 3 IV.3.) a permis par la suite, lors des analyses, de classer chacun des événements vocaux relevés : soit à l'intérieur d'un tour de parole du sujet, noté « 1 » lorsque ce dernier est en train de produire de la parole, soit noté « 0 » dans les autres cas, c'est-à-dire en dehors d'un tour de parole du sujet (voir Chapitre 2 I.4.1. et Chapitre 3 II.2.4.), sauf cas particuliers relevés ci-dessous.

Les tâches correspondant à des tours de parole (« 1 ») sont « pendant commentaires », « pendant page suivante », « pendant réponse » et « pendant prononciation ».

Les tâches notées « 0 » (en dehors du tour de parole du sujet) sont « pendant lecture », « entre prononciation », « avant réponse », « après réponse », « après page suivante » et « après commentaires ».

Selon les définitions de la pragmatique, qui reviennent à déterminer si c'est au sujet ou non de parler, certains de nos cas serait alors problématiques, et probablement classés pendant le tour de parole (« 1 » alors que nous les avons notés « 0 ») puisque que le système reste, lors de ces tâches, en attente d'une réponse / commande de la part du sujet . Ces cas particuliers concernent les tâches :

- « entre prononciation » et « avant réponse » ;
- « après commentaires », tâche qui sous-entend « avant page suivante », que nous avons notée « 0 », quand bien même tant que le sujet n'a pas dit « page suivante », c'est toujours « son tour de parler ».

Concrètement, lors de l'étiquetage des événements vocaux de notre corpus, nous avons inclus l'événement à l'intérieur du label de tâche noté « 1 » lorsque son occurrence se trouvait à moins d'une demi seconde avant ou après la prise de parole.

### III.1.2. Précisions sur les relations entre parole et événements vocaux, et sur leur notation

Nous avons classé chacun des événements vocaux en fonction de l'organisation du temps de la communication.

Ces relations entretenues avec la parole peuvent être de différents ordres et ont ainsi été notés à différents niveaux : soit dans l'étiquette elle-même de l'occurrence de l'événement vocal, soit dans une *track* indépendante du fichier d'annotation (Voir fichier de spécification ANVIL en Annexe 8), et dont la relation avec le ou les événements vocaux en question a été étudiée lors des analyses.

Au sein même de l'étiquette de l'occurrence, nous avons directement noté la position de celle-ci par rapport à la production de parole du sujet, en utilisant le code suivant :

- « 1 » lorsqu'ils apparaissaient pendant une prise de parole et non coupés du flux ;
- « 0 » en dehors de toute prise de parole, c'est-à-dire à plus d'une seconde de l'une d'entre elle ;
- « B » (pour *Before*) moins d'une seconde avant une prise de parole ;
- « A » (pour *After*) moins d'une seconde après une prise de parole ;
- « ~ » lorsqu'ils apparaissaient pendant une prise de parole, mais en dehors du flux de parole, c'est-à-dire coupés du flux par une pause de moins d'une seconde. En quelque sorte, « ~ » est une combinaison de « A » et « B ».

Par prise de parole, nous avons considéré toute intervention vocale, telle un commentaire libre du sujet, une réponse à un stimulus, la production d'une voyelle (à la demande du logiciel), ou encore la production de la commande vocale « page suivante ».

Nous avons aussi considéré comme prise de parole la production d'une interjection (autre que celle dont l'occurrence est considéré). D'un point de vue concret, cela signifie que lorsque deux interjections se suivaient à moins d'une seconde d'intervalle dans notre corpus, la première était notée « B » et la seconde « A » à moins que l'une d'entre elle se trouve à moins d'une seconde d'une autre prise de parole, et soit alors notée « ~ ». De la même manière, un événement vocal autre qu'une interjection (voir plus loin pour la distinction), quant à lui non considéré comme prise de parole, mais dont l'occurrence se trouve à moins d'une seconde d'une interjection, sera alors noté « A » ou « B » (selon le cas).

Quant aux moments où le sujet a lu ou parlé à voix basse (d'ailleurs il s'agit le plus souvent de parole non compréhensible), ils n'ont pas été considérés comme prise de parole. Il s'agit d'un cas fréquent pour le sujet M_J, mais très peu présent chez les autres sujets.

### III.1.3. Choix des sujets

Pour cette partie de nos recherches portant sur les événements vocaux, six sujets ont été étudiés. En effet, encore plus que pour les expressions gestuelles/faciales, les événements vocaux et la manière de les utiliser sont très variables selon les sujets. C'est pourquoi afin d'avoir plus de données vocales sur lesquelles travailler, et pour pouvoir comparer, c'est-à-dire trouver des régularités et des différences, entre le comportement des différents sujets, nous avons étendu l'étiquetage puis l'analyse des événements vocaux à quatre autres sujets. Ces six sujets ont été retenus pour la variabilité de leurs réactions affectives et cognitives aux inductions, et pour la variabilité de leur personnalité perçue naïvement (d'introverti à extraverti, voire exubérant, selon des auditeurs naïfs). Il s'agit de trois hommes et trois femmes (notés respectivement M_ et F_ (pour « masculin » et « féminin ») suivi de la première lettre de leur prénom, afin de préserver leur anonymat. Ils sont d'âge, de niveau d'éducation, et de familiarité aux technologies de type IHM variables. Le Tableau 4 résume cette variation intrinsèque inter-sujets.

Tableau 4: les sujets, leurs caractéristiques et leurs ressentis par phase
pour la description des différentes phases au niveau du scénario, se reporter au Chapitre 3 II.2.2. ; quant à l'évolution globale des états des sujets, il s'agit des mots ou expressions qui revenaient le plus souvent dans les auto-annotations du sujet en question

| Sujet | Sexe | Tranche d'âge | Niveau d'éducation | Familiarité avec les technologies informatiques | introverti / extraverti (jugement par des auditeurs naïfs) | remet en question : | Evolution globale des états | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Phase 1 | Phase 2 | Phase 3 | Phase 4 |
| F_S | féminin | 20-25 | élevé | élevé | introvertie | la machine | stress et concentration | confiance en soi, surprise | concentration, déception, réponse au hasard | déception, "rire jaune", tentative d'explication |
| F_T | féminin | 20-25 | élevé | élevé | extravertie | soi-même | perplexité, concentration, stress | malaise, déception, stress, rassurée | calme, déception, perplexité, stress | hésitation, stress et même angoisse |
| F_M | féminin | 20-25 | élevé | élevé | introvertie | la machine | sûre de soi | doute | doute et déception | agacement, amusement |
| M_J | masculin | 40-45 | moyen | moyen | introverti | la machine | concentration, intrigué | inquiètude | doute | indifférence, irritation |
| M_R | masculin | 30-35 | moyen | moyen | extraverti | soi-même | concentration, ennui | concentration, ennui | surprise, agacement, concentration | ennui, impatience de finir, agacement |
| M_N | masculin | 20-25 | élevé | moyen | introverti | soi-même | doute, pas sûr de soi | doute, gêne | doute, agacement | agacement, déconcentration |

## *III.2. Classement des événements vocaux : critères et autres paramètres*

### III.2.1. Généralités sur le classement

Dans notre corpus, un grand nombre d'événements vocaux ont été étiquetés. Comme le laissait présager l'état de l'art, ces événements peuvent être très différents de par leurs paramètres acoustiques et articulatoires.

Pour étiqueter ces événements vocaux, nous avons d'abord distinguer :

- les interjections : éléments pré-lexicaux construits avec des phonèmes de la langue et dont une transcription conventionnelle existe la plupart du temps dans les dictionnaires de la langue ;
- des autres événements vocaux -que nous avons globalement nommés « bruits de bouche » pour faire court- : construits de *phones vs. non-phones* (*e.g.* bruits d'occlusions, de respiration, clicks, raclements de gorge, gémissement, etc.).

La signification socialement acceptée devient ainsi un critère pour la classification lorsque l'événement vocal peut être transcrit avec une orthographe conventionnelle. Nous pouvons même observer que certains événements vocaux peuvent être produit avec des variations phonologiques. Par exemple en français, l'interjection

orthographisé « pfff » peut être produite articulatoirement [pf:] ou avec un *trill* bilabial.

Ensuite, nous avons classé les interjections selon des critères phonétiques / articulatoires : phonème vocalique (étiqueté « V », *e.g.* [ø:] « euh »), phonème consonantique (étiqueté « C », *e.g.* [m:] « mmh ») ou combinaison de phonèmes vocaliques et consonantiques (étiqueté « CV », *e.g.* [bø:] « beuh », et « VC », *e.g.* [ø:m:] « euh mmh »). Lorsqu'une interjection est composé de plus d'un phonème vocalique ou consonantique, nous l'avons étiqueté « Comb » (pour « combinaison »), *e.g.* [bɛ̃m:] « ben mmh », [ula] « ouh là ». Nous avons donc étiqueté des événements vocaux complexes, ainsi que des plus simples.

### III.2.2. Impression articulatoire

Nous avons décrit les bruits de bouche selon plusieurs paramètres, chacun faisant l'objet d'un champ (attribut d'ANVIL) particulier dans la fenêtre d'annotation correspondante à la *track* « bruit de bouche » (voir le fichier de spécification d'ANVIL Annexe 8)

Il est fondamental de rappeler que nous ne disposons en aucun cas de mesure articulatoire de la production[96]. Les critères qui suivent sont donc « impressionnistes », c'est-à-dire d'observation auditive et visuelle, éventuellement confirmée par des analyses du signal acoustique. Notons que, précisément pour ces sons qui n'entrent pas tous dans les formes connues de la phonétique, les analyses de signal de parole sont sûrement à adapter.

Les critères retenus pour les bruits de bouche sont :

- leur nature articulatoire : relâchement d'articulateur, click, occlusion, occlusion ingressive suivi d'une inspiration (« occlusion ingressive_inspiration »), inspiration simple, expiration simple, relâche sa respiration, expiration brutale, friction, gémissement, raclement de gorge, déglutit, interaction entre langue et lèvres provoquant un bruit de « mouillé » (« interaction langue-lèvres mouillées »), et « autre »[97] (*cf.* partie III.3. pour la manière dont cette liste a été établie) ;

- l'événement vocal qui le suit directement, éventuellement : les étiquettes possibles sont les mêmes que pour la nature articulatoire, avec en plus une étiquette « non lieu ». Pour que deux événements soient étiquetés comme un seul élément (et non

[96] Nous ne disposons pas même de signal EGG, capturé pour plusieurs sujets de SoundTeacher, mais pour aucun des 6 sujets que nous avons retenus.

[97] Notons que certains types vont être toujours associés aux mêmes paramètres, alors que d'autres vont pouvoir variés et ainsi être associés à différents paramètres selon la forme de leur occurrence (voir Chapitre 6 I.2.).

deux éléments juxtaposés, les critères sont qu'ils doivent être produits dans un même flux d'air (égressif ou ingressif) et qu'aucune pause entre eux ne soit perceptible et ni étiquetable ;

- leur lieu d'articulation s'il était pertinent de l'indiquer ;

- leur qualité de voix : de la même manière que pour les interjections, avec en plus l'étiquette « tremblante » (voir une précision sur cette notion, et sur la manière de la noter pour les interjection dans la partie suivante - III.2.3. -) ;

- le type d'intensité et de contrôle : son abrupte / intense, son laxe, ou encore ni l'un ni l'autre, car contrôlé volontairement ou non ;

- 5 étiquettes de type oui / non : « flux d'air nasal », « bruit de mouillé », « voisement », et aussi « visible » et « peu audible » en tant qu'étiquettes subjectives à valeur indicative.

Précisons la manière de compter les événements vocaux lors des analyses, en regard de la présence du paramètre « élément qui suit » : lors du recensement global des bruits de bouche, chaque événement, qu'il soit en première ou en deuxième position, a été compté pour le recensement (Chapitre 6 I.1.) ; dans toutes les autres analyses, les deux événements qui se suivent sont considérés comme une seule entité, puisque tous les autres paramètres étiquetés correspondent aux couples d'événements. Concrètement, dans les tableaux de données extraits, chaque ligne correspond à un événement, ou un couple d'événements, associés à différents paramètres intrinsèques ou situationnels. Au niveau des bruits de bouche étiquetés, cela explique la différence entre leur nombre dans le recensement global (2521), et dans les analyses de leurs liens aux différents paramètres (2358, sachant que nous avons noté 163 événements « double », c'est-à-dire dont le paramètre « élément qui suit » n'est pas « non lieu »).

Par ailleurs, pour les premières analyses portant sur l'organisation temporelle des bruits de bouche, nous avons particulièrement utilisé deux distinctions fondées sur des paramètres articulatoires pour classer les bruits de bouche :

- s'ils sont produits par le sujet pendant un un flux d'air ingressif *vs.* égressif (*cf.* (Eklund, 2008), pour plus de précisions concernant la parole produite dans un flux d'air ingressif) ;

- s'ils sont produits avec un flux d'air continu (*e.g.* grande inspiration, gémissement) *vs.* avec un flux d'air gêné, ou bloqué au moins une fois (y compris occlusion glottale, *e.g.* click, occlusive bilabiale). Ce critère est particulièrement intéressant car il implique une tension suivie d'un relâchement du sujet, dont nous reparlerons Chapitre 7.

Les bruits de bouche de type « mouillé », liés soit à une interaction entre langue et lèvres, soit à une déglutition du sujet, ne peuvent être classés en utilisant ces deux distinctions. Ils ont donc été classés comme types à part lors des analyses fondées sur ces oppositions.

### III.2.3. Qualité de son / qualité de voix / qualité de voyelle

Nous utilisons pour qualifier les événements vocaux des critères qui sont pour la plupart empruntés à la phonétique (*i.e.* qui peuvent caractériser des sons des langues). Pourtant, la particularité de ces sons est qu'ils n'entrent pas dans les sons du français, sauf pour les interjections, ni, pour beaucoup, dans les sons des langues du monde.

La qualité de voix est en soi un problème complexe (*cf.* par exemple Audibert, 2008, pour un état de l'art) : peu est connu de la production glottique, encore moins de la production supra-glottique, pour laquelle il n'existe que très peu d'outils corrects d'analyse du signal. La qualité de voix a été étudiée pour les voix produisant les sons d'une langue. Ces sons portent alors une information par ces variations acoustiques de qualité de voix, sans détruire, en particulier, le substrat phonologique de la qualité des voyelles. La qualité de voix est ainsi une variation autour de la qualité des sons du langage, et autour de la qualité des voyelles en particulier. Or, les bruits de bouche ne sont pas, ou pas encore, des sons du système phonologique de la double articulation. Par conséquent, même si la voix s'adresse à tous les sons produits par le conduit vocal, et pas seulement ceux de la parole, il serait certainement inapproprié de parler directement de qualité de voix pour les bruits de bouche. Cependant, de la même manière que nous utilisons autant que possible les critères articulatoires utilisés pour les sons du langage pour qualifier les bruits de bouche, nous utiliserons autant que possible, les mêmes critères que la voix parlante.

Par contre, et il est essentiel de le noter, à aucun moment nous signifions par là que la qualité sonore se qualifie strictement par les critères que nous avons décrits plus haut. Bien au contraire nous mettons ces critères, impression articulatoire et impression de qualité de son, à un même niveau de description. En effet, c'est ce qui permettra ultérieurement de poser la question d'une continuité émergente entre une qualité de son, non phonétique mais directement signifiante, et la qualité de voix sur les segments de la double articulation.

Ainsi, nous empruntons d'une part à la description des sons phonologiques, et d'autre part à la qualité de voix, qui dans la perspective linguistique traditionnelle n'entre pas dans la double articulation (*i.e.* dans l'accès lexical). Malgré cela, ces deux voies sont, pour nous et sur les objets décrits ici, une seule et même entité de description des formes. Ainsi, nous laissons toute liberté à des travaux ultérieurs de ramener à de la

qualité supra-glottique certains critères articulatoires, et de ramener à des critères articulatoires certains contrôles de qualité de voix.

Concernant notre étiquetage, nous avons noté de manière impressionniste les qualités glottiques : murmuré, chuchoté, soupiré, voix craquée (*creaky voice*), et nasalisé (noté comme qualité de voix lorsque les phonèmes produits ne nous semblaient pas intrinsèquement des sons nasaux). Ces éléments comportant de la qualité de voix ont été traités comme X-variante en terme de type dans les analyses par type prenant en compte ce paramètre (*e.g.* « V-variante », « CV-variante », etc.).

## *III.3. Les difficultés d'une méthodologie empirique et les questions immédiates découlant de l'étiquetage*

La liste des types de bruits de bouche utilisées dans les analyses (indiquée p.206) est une liste établie à partir du classement d'observations décrites « naïvement » (c'est-à-dire sans se référer à une théorie préalable) par deux linguistes de formation (Loyau et Vanpé). Cette liste a ainsi été ajustée et modifiée à maintes reprises, en particulier à chaque nouveau sujet étiqueté. En effet, étiqueter un nouveau sujet nous a souvent permis de nous rendre compte de la pertinence d'un paramètre qui n'avait pas été pris en compte jusque là. Ainsi, ce sont les similitudes mais également les dissemblances des comportements des sujets qui ont guidé la classification et les paramètres relevés.

D'un point de vue pratique, cela signifie que l'étiquetage des 6 sujets a également été reprise plusieurs fois, de manière à ce que les sujets soient tous étiquetés sur la base de la même grille et avec une homogénéité perceptive. En effet, un même type d'événements vocaux, va, malgré son homogénéité articulatoire, pouvoir varier au sein de ses occurrences, et d'autant plus entre les sujets. Nous approfondirons ce point Chapitre 6 III.

Par ailleurs, le problème de la normalisation de la terminologie s'est posé, en particulier en ce qui concerne les bruits de bouche et surtout lors de l'étiquetage de type « description libre » utilisée lors de la première annotation des sujets concernant les événements vocaux. Avant d'avoir la grille d'étiquetage finale, il a fallu homogénéiser au fur et à mesure les noms utilisés pour les descriptions des divers événements rencontrés, de manière à faciliter par la suite leur traitement automatique : nommer les événements similaires avec exactement le même nom, et avec un nom explicite, précis et objectif - c'est-à-dire selon des critères acoustiques / articulatoires- (*e.g.* ne pas utiliser « rire », beaucoup trop vaste et subjectif) ; mettre la liste des variables observées entre parenthèse, toujours écrite de la même manière et dans le même ordre ; utiliser le signe « + » dans le cas de succession d'événements ; etc.

Nous nous sommes également confrontés à la problématique de la classification, à savoir quels sont les événements similaires, ou en quoi les événements sont différents. Les questions liées à cette classification ont été résolues par les critères acoustiques / articulatoires fixés (type de flux d'air, contraintes articulatoires exercées sur lui, etc. voir ), testés en ayant recours au couple imitation - proprioception.

Ainsi, à titre d'exemple, les distinctions entre les différents bruits de bouche courts et produits avec un flux d'air ingressif et bloqué, a été remis en question au cours de l'étiquetage. En effet, ces événements proches en termes de description globale, mais apparaissant intuitivement différents perceptivement, nous ont amenés à introduire un nouveau critère de classification de type contrôle articulatoire, qui les a séparés en trois classes :

- « relâchement articulateur » : bruit la plupart du temps « mouillé », souvent peu audible, conséquent d'un relâchement musculaire d'articulateur(s) ;
- « occlusion », en l'occurence ingressive, déterminée selon la définition articulatoire classique, à savoir une interruption totale du passage de l'air par les cavités orales, le son résultant ensuite de la libération brutale de l'air interne ;
- « clicks », également selon la définition articulatoire classique, c'est-à-dire deux occlusions, une à l'avant du conduit vocal, l'autre à l'arrière, puis la création d'une dépression buccale par l'abaissement de la langue, le son résultant enfin du relâchement brutale de l'occlusion la plus en avant. Dans la majorité des cas, les clicks sont très audibles et produits de manière « abrupte / intense ».

Un élément nous conforte dans la réalisation de cette distinction : dans notre corpus, avant une inspiration, l'air bloqué peut être la conséquence soit d'une occlusion (ingressive en l'occurrence), soit d'un click, mais jamais d'un « relâchement articulateur ». Une autre classe proche mais indépendante a également été distinguée : « occlusion ingressive_inspiration », qui apparaît comme le parallèle de type ingressif de « relâche sa respiration », puisque l'air est dans les deux cas bloqué au début de l'événement (en général au niveau glottal pour « relâche sa respiration »), puis libre après le relâchement et dans sa continuité.

Les critères mêmes de classification des événements vocaux ont donc été à déterminer et des cas restent problématiques à la suite de l'étiquetage :

- certains bruits de bouche sont à la limite de l'interjection au niveau perceptif. Il s'agit souvent de cas présentant du voisement, une « qualité de son », un flux d'air gêné, et / ou une prise de contrôle. C'est en particulier à ces événements-là que nous allons nous intéresser lors d'analyses acoustiques. Ces analyses nous permettront de préciser les différences existantes entre ces bruits de bouche particuliers et les interjections, et

ainsi de mieux appréhender les paramètres contrôlés, volontairement ou involontairement.

- d'autres bruits de bouche sont presque inaudibles. Sachant que l'étiquetage a été réalisé de manière audio-visuelle, la question est alors le rôle du visuel dans la perception de l'audio, et le type d'information qu'il apporte éventuellement.

- puisqu'il apparaît une adaptation et un apprentissage sur le sujet, et qu'il existe des événements vocaux idiosyncrasiques, la valeur communicative de ces derniers reste à connaître : ces indices peuvent-ils devenir signaux lorsque nous connaissons la personne ? Comment l'information apportée par ces événements évolue-t-elle lors de « l'apprentissage » du sujet ? Le gain d'information potentiellement apporté par l'« apprentissage » du sujet, concerne-t-il la culture, la personnalité du sujet, ou ses états de *FoT* ?

Nous allons tenter d'apporter des éléments de réponse à ces interrogations posées à la suite de l'étiquetage, en ayant recours à des analyses acoustiques, des évaluations perceptives (voir Chapitre 7 II.2.) et des analyses de répartition temporelle en fonction de facteurs situationnels et intrinsèques.

# IV. Résumé

Dans la littérature, les événements vocaux ont été décrits différemment selon les domaines ou les objectifs de recherche : de la rhétorique à la pragmatique, en passant par l'étude de l'évolution du langage. De ce fait, ils n'ont pas encore été vraiment décrits et organisés longitudinalement, dans une même perspective que celle de leur étude phonétique via leur fonctions. Concernant les interjections en particulier, c'est-à-dire les non-mots les plus proches d'un accès lexical structurable en morpho-syntaxe, des débats se sont focalisés sur la question de leur appartenance à une partie du discours (*i.e.* de leur nature linguistique).

Les technologies de la communication parlée et non verbale ont quant à elles mis en évidence l'importance de ces « détails » phonétiques. Étant donné qu'elles restreignent fortement les situations de communication, elles ne contrôlent pas certaines parties du comportement de l'humain virtuel. Ainsi, elles démontrent implicitement que les compétences communicatives de l'humain ne sont pas réductibles. Par exemple les tout premiers agents virtuels ne s'impliquaient pas dans leur énonciation, et ne communiquaient rien en dehors de leurs tours de parole. Ils ont été des démonstrations implicites de la nature informative de l'expressivité (*cf.* les travaux extrêmement nombreux depuis Picard, 1997, en particulier Pelachaud et collègues) ainsi que de la nature informative des gestes et sons produits en dehors des tours de parole *(e.g.* Poggi, Pelachaud, & Magno Caldognetto, 2004, ou Heylen, Nijholt, & Poel, 2007).

Quoi qu'il en soit, l'enjeu théorique de l'étude des événements vocaux est aussi important que les enjeux technologiques. Cependant, il nécessiterait d'élargir l'objet d'étude, souvent restreint aux interjections ou à d'autres événements ayant une fonction précise, à l'ensemble des événements vocaux rencontrés en situation spontanée. Ainsi, nous considérerons et analyserons l'ensemble des événements et micro-événements vocaux, dans « tri » au préalable.

Bien que nous distinguons deux catégories dans nos analyses, nous cherchons à classer nos événements vocaux sur un continuum qui va du bruit de bouche, étant éventuellement un indice involontaire, aux items pré-lexicaux construits par une qualité sonore qui fait déjà partie du système phonologique de la langue (*i.e.* une qualité de voyelle. Entre ces deux extrêmes se trouvent des sons qui ne pourraient des phones (sons phonétiques) d'une autre langue que celle du sujet, et des sons que nous essaierons de décrire par des indices phonétiques.

# CHAPITRE 6 : LES ÉVÉNEMENTS VOCAUX AU SEIN D'UNE COMMUNICATION TEMPORELLEMENT SITUÉE

Nous avons montré dans le chapitre précédent l'intérêt d'étudier les événements vocaux de manière globale dans un premier temps, en nous attardant particulièrement sur les paramètres communicationnels de leur production.

Sont-ils produits en fonction d'événements organisant la tâche et la communication ? Leur « densité » est-elle organisée temporellement ? Est-elle dépendante des productions de parole, des états affectifs, de la tâche, du sujet ? Les observations liées à ces questions sont-elles toujours valables si nous considérons indépendamment les différents types d'événements vocaux ? Et, au préalable, quels sont les événements vocaux les plus récurrents dans nos données ?

Répondre à ces questions permettrait d'éclairer l'organisation temporelle de ces événements. Il serait alors possible de les situer au sein des différentes organisations temporelles (local et global) d'une part, et de mieux déterminer de quels niveaux temporels relèvent ces événements d'autre part.

Pour cela nous allons d'abord effectuer un recensement global des différents type d'événements vocaux présents dans nos données. Puis nous appréhenderons leur organisation temporelle à travers l'étude des liens existants entre les types d'événements vocaux et certains des paramètres situationnels que nous avons étiquetés. Cela nous amènera à préciser les variations inter-individuelles existant dans le comportement vocal des sujets, et ainsi à nous intéresser au rôle des motifs temporels dans la caractérisation comportemental des personnes. Enfin, nous réfléchirons sur les perspectives apportées par les motifs temporels, la problématique liée à leur étude et sur les méthodologies qu'il est possible de mettre en œuvre, en particulier par le biais de l'étude de la rythmicité, issue d'une approche éthologique.

Rappelons que les analyses effectuées dans ce travail restent descriptives et limitées à nos données, et que ces dernières sont issues d'une interaction située ayant ses particularités. Il reste donc à déterminer si les observations qui suivent sont statistiquement significatives, puis dans ce dernier cas, à établir dans quelle mesure elles pourraient avoir des caractères génériques [98].

[98] Dans tous les cas, la simulation des sujets dans des tâches IHM analogues reste une application directe.

# I. Recensement global des événements vocaux

## *I.1. Des événements de nature variée*

Pour les 6 sujets et les 3h45 de vidéo traitées correspondantes, nous avons étiqueté un total de 3011 événements vocaux : 84% (2521) de bruits de bouche et 16% (490) d'interjections (rappelons que notre définition de l'interjection est quelque peu arbitraire et que les événements vocaux s'organisent sur un continuum). Un inventaire phonétique des bruits de bouche (voir Tableau 5) nous montre que 76,1% des bruits de bouche se répartissent dans seulement 5 des 13 catégories que nous avons définies : « inspiration », « relâchement articulateur », « expiration », « occlusion ingressive_inspiration » et « déglutit ». Il est à noter que ces cinq catégories concernent toutes des bruits de bouche dont la production requiert un contrôle articulatoirement peu complexe (par rapport aux contrôles de cibles phonologiques).

Tableau 5: Inventaire phonétique des bruits de bouche,
des plus fréquents aux moins fréquents.

| Type | Nombre d'occurrences | Pourcentage du total |
|---|---|---|
| inspiration | 577 | 22,9% |
| relâchement articulateur | 477 | 18,9% |
| expiration | 393 | 15,6% |
| occlusion ingressive_inspiration | 251 | 10,0% |
| déglutit | 220 | 8,7% |
| occlusion | 151 | 6,0% |
| relâche sa respiration | 127 | 5,0% |
| expiration brutale | 126 | 5,0% |
| interaction langue-levres mouillees | 92 | 3,6% |
| click | 76 | 3,0% |
| friction | 13 | 0,5% |
| raclement de gorge | 10 | 0,4% |
| gémissement | 8 | 0,3% |
| **Total** | **2521** | 100,0% |

Quant aux interjections, les plus fréquentes sont celles de type vocalique pur (V, 64,5% de l'ensemble des interjections), que ce soit avec ou sans l'utilisation de qualité de voix particulière (voir Chapitre 5 III.2. pour un rappel des types).

Sur l'ensemble des interjections relevées, 74% (361 sur 490) n'ont pas de qualité de voix particulière, et sont donc notés « modal ».

Tableau 6: Inventaire des interjections relevées
par ordre de fréquence en fonction du type phonétique.

| Type | Nombre d'occurrences | Pourcentage du total |
|---|---|---|
| V (modal) | 226 | 46,1% |
| V-variante | 90 | 18,4% |
| **Total V** | **316** | **64,5%** |
| VC (modal) | 46 | 9,4% |
| VC-variante | 12 | 2,4% |
| **Total VC** | **58** | **11,8%** |
| CV (modal) | 45 | 9,2% |
| CV-variante | 7 | 1,4% |
| **Total CV** | **52** | **10,6%** |
| C (modal) | 23 | 4,7% |
| C-variante | 11 | 2,2% |
| **Total C** | **34** | **6,9%** |
| Comb (modal) | 21 | 4,3% |
| Comb-variante | 9 | 1,8% |
| **Total Comb** | **30** | **6,1%** |
| **Total** | **490** | **100,0%** |

Nous reviendrons plus en détail sur l'utilisation des différentes qualités de voix en fonction des types phonétiques d'interjection Chapitre 7 I.

À travers ces deux tableaux, ainsi que dans le Tableau 7 récapitulatif, nous observons que les événements vocaux ont des fréquences très différentes selon leur type. Ainsi, les trois types d'événements les plus fréquents (inspiration, relâchement d'un articulateur et expiration) sont des bruits de bouche, et sont près de la moitié des événements. A contrario, les 9 catégories (soit la moitié) les moins fréquentes couvrent seulement 12,4% des événements relevés. Ces catégories peu fréquentes correspondent aux : interactions langue-lèvres mouillées ; clicks ; interjections VC, CV, C et « Comb » ; frictions ; raclements de gorge ; et gémissements.

De plus, seules les interjections vocaliques représentent plus de 10% des événements étiquetés à elles-seules.

Tableau 7: Inventaire global des événements vocaux par type
et par ordre décroissant de fréquence.

| Type | Nombre d'occurrences | Pourcentage du total |
|---|---|---|
| inspiration | 577 | 19,2% |
| relâchement articulateur | 477 | 15,8% |
| expiration | 393 | 13,1% |
| V | 316 | 10,5% |
| occlusion ingressive_inspiration | 251 | 8,3% |
| déglutit | 220 | 7,3% |
| occlusion | 151 | 5,0% |
| relâche sa respiration | 127 | 4,2% |
| expiration brutale | 126 | 4,2% |
| interaction langue-levres mouillees | 92 | 3,1% |
| click | 76 | 2,5% |
| VC | 58 | 1,9% |
| CV | 52 | 1,7% |
| C | 34 | 1,1% |
| Comb | 30 | 1,0% |
| friction | 13 | 0,4% |
| raclement de gorge | 10 | 0,3% |
| gémissement | 8 | 0,3% |
| **Total** | **3011** | **100%** |

## *I.2. Les relations entre types d'événements et flux d'air*

Les différents types articulatoires/acoustiques d'éléments relevés peuvent être de nature plus ou moins complexe. Ceci est notamment observable par les différents paramètres qu'un même type d'élément est susceptible de prendre. En effet, alors que certains types vont être toujours associés aux mêmes paramètres, d'autres vont pouvoir varier et ainsi être associés à différents paramètres selon la forme de leur occurrence. Nous traiterons de ces derniers cas Chapitre 7 partie I.

Les liens constants, ou quasi-constants, sont souvent dus à la définition même de la catégorie. Par exemple une inspiration sera, par définition, produite dans un flux d'air ingressif et continu, alors qu'un relâchement de la respiration sera produit dans un flux d'air égressif et bloqué.

« Expiration », « expiration brutale », « gémissement », « raclement de gorge » et « relâche sa respiration » sont presque uniquement produits dans un flux d'air égressif. Au contraire, « click », « inspiration », « occlusion ingressive_inspiration » et « relâchement articulateur » le sont uniquement dans un flux d'air ingressif (voir le tableau détaillé Annexe 9). Quant aux catégories « frictions » et « occlusions », elles ne sont pas exclusivement produites dans un flux d'air particulier : un quart des frictions sont égressives et 75% ingressives ; et 55% des occlusions sont égressives, vs. 45% sont ingressives.

Tableau 8: Nombre et proportion de bruits de bouche selon leur type de flux d'air.

| Flux d'air | | Total | % du total d'égressif/ ingressif | % du total de BB |
|---|---|---|---|---|
| égressif | bloqué | 239 | 32% | 11% |
| | gêné | 69 | 9% | 3% |
| | continu | 437 | 59% | 21% |
| ingressif | bloqué | 861 | 63% | 41% |
| | gêné | 52 | 4% | 2% |
| | continu | 461 | 34% | 22% |

Regrouper les différents bruits de bouche selon leur flux d'air fait apparaître que 65% (41%+2%+22%) des bruits de bouche relevés sont produits dans un flux d'air ingressif (Tableau 8). Plus précisément, la majorité (63% d'entre eux, et 41% du total) ont ce flux d'air bloqué lors de la production. À l'inverse, la majorité (59%) des bruits de bouche égressif sont produits dans un flux d'air continu. En parallèle, 78% (861/1100) des bruits de bouche « bloqués » sont produits dans un flux d'air ingressif, alors que les bruits de bouche « continus » sont bien répartis au niveau du flux d'air dans lequel ils sont produits (ingressif pour 51%, et égressif pour 49%). Par ailleurs, seulement 5% des bruits de bouche sont produits avec un flux d'air « gêné ».

Nous reviendrons Chapitre 7 I. sur les occurrences de bruits de bouche produits avec les différents flux d'air. Nous développerons alors le lien existant entre ce dernier critère et les paramètres de voisement et de qualité de voix.

Notons par ailleurs que les interjections relevées sont toujours produites dans un flux d'air égressif. Rappelons que nos données ont été recueillies avec des sujets français en interaction personne-machine. Cela ne signifie donc pas qu'il n'existe pas d'interjections produites dans un flux d'air ingressif dans d'autres langues, ou même dans d'autres contextes (*e.g.* l'interjection française [h$^{n}$ã] pouvant exprimer la surprise est le souvent ingressive)(*cf.* Eklund, 2008 pour plus de détails).

Seuls les événements vocaux « mouillés », c'est-à-dire « déglutit » et « langue-lèvres mouillées » ne sont pas rattachés dans l'absolu à un flux d'air égressif ou ingressif particulier.

Dans les cas de liens constants, étudier les aberrations relevées et leur nature permet d'éclairer certains facteurs de variations inter- et intra-individuelles. En effet, cela attire l'attention sur la façon dont un événement peut être modifié par son environnement immédiat (en particulier temporel). Ainsi, c'est souvent lorsque l'événement est suivi par un autre élément, que ses paramètres définitionnels sont modifiés. En particulier, « déglutit » et « interaction langue-lèvres mouillées » (dont les paramètres « flux d'air » sont habituellement notés « non lieu ») sont produits dans un flux d'air particulier lorsqu'ils sont couplés à un autre événement (*e.g.* « déglutit + relâche sa respiration » implique une flux d'air égressif). C'est le cas de 52 « déglutit » sur 219 (24%), et de 14 « interaction langue-lèvres mouillées » sur 86 (16%).

# II. Des liens complexes entre types et paramètres situationnels associés

Différents paramètres de nature situationnelle ont été notés lors de l'étiquetage. Pour un rappel concernant le scénario et son étiquetage, et l'étiquetage des paramètres temporels, voir Chapitre 3 IV.3. Ces paramètres permettent d'analyser les événements vocaux :

- soit dans une organisation temporelle globale, c'est-à-dire inscrite dans l'organisation générale du paradigme du scenario de l'interaction. Dans les présentes analyses, nous prendrons en considération les paramètres « Phase », et « Tâche Globale » ;
- soit dans une organisation temporelle locale, à une granularité inférieure, c'est-à-dire en lien avec les événements communicatifs syntagmatiquement proches. Sont alors considérés les paramètres « élément qui suit », le lien à la prise de parole, et la tâche dans laquelle le sujet se trouve.

Notons que les distinctions entre organisation temporelle locale et globale, et entre niveau événementiel et niveau communicatif, ne sont pas priori dépendantes : par exemple au sein de l'organisation temporelle locale se trouvent à la fois des paramètres liés, par définition, au temps événementiel (*e.g.* la tâche dans laquelle le sujet se trouve), et au temps communicatif (*e.g.* le lien à la prise de parole).

Les objectifs à moyen et long terme de cette analyse des différentes organisations temporelles des événements vocaux sont de modéliser pourquoi, comment et quand la modalité acoustique est utilisée lors d'une tâche cognitive. Nous cherchons également à comprendre dans quels cas elle est liée ou non à l'organisation de l'interaction. Par exemple il a été montré que la forme d'un rire est révélatrice de son contexte pragmatique d'apparition (Loyau, 2007 ; Petridis, 2008), et d'éléments d'actes du langage (Searle, 1979 ; Vanderveken, 1990).

Dans un premier temps, outre l'étude de l'influence des différents paramètres situationnels, il s'agit d'analyser si le comportement de chacun des types particuliers d'événements vocaux, pris indépendamment, va dans le même sens que les tendances globales de comportements des événements.

## *II.1. Lien à la phase du scénario*

Globalement, les événements vocaux sont répartis équitablement au cours des différentes phases du scénario (de 24 à 26% selon les phases -Tableau 9-), avec 712 événements vocaux par phase en moyenne, et un coefficient de variation de 0,02 (Tableau 10). Rappelons que la phase 1 correspond à une tâche de familiarisation, avec renforcement positif ; la phase 2 à un renforcement positif fort ; la phase 3 aux premières inductions négatives ; la phase 4 à une tentative de déstabilisation du sujet, avec des inductions négatives fortes.

Tableau 9: Pourcentage d'événements vocaux par phase selon leur catégorie

| Phase | Bruits de bouche | Interjections | Total |
|---|---|---|---|
| 1 | 26% | 21% | **25%** |
| 2 | 23% | 30% | **24%** |
| 3 | 26% | 22% | **25%** |
| 4 | 25% | 28% | **26%** |
| Total | 100% | 100% | **100%** |

Une observation intéressante apparaît en considérant le pourcentage d'interjections en fonction du nombre total d'événements vocaux selon la phase : ce pourcentage est plus élevé dans les phases 2 et 4 (avec respectivement 21 et 19%), que dans les phases 1 et 3 (14 et 15%)[99]. En parallèle, les bruits de bouche sont globalement bien répartis dans les quatre phase (de 23 à 26%) : moyenne de 589,5 bruits de bouche par phase, avec un coefficient de variation de 0,05 (Tableau 10).

Tableau 10: Nombre d'événements vocaux par phase, et moyenne, écart-type et coefficient de variation global selon leur catégorie.

| Phase | BB | Interj | Total |
|---|---|---|---|
| 1 | 603 | 101 | **704** |
| 2 | 547 | 145 | **692** |
| 3 | 617 | 108 | **725** |
| 4 | 591 | 136 | **727** |
| **Total** | **2358** | **490** | **2848** |
| Moyenne | **589,5** | **122,5** | **712** |
| Ecart-type | 30,26 | 21,30 | 16,91 |
| Coef. de variation | 0,05 | 0,17 | 0,02 |

Plus précisément, si nous considérons indépendamment les différents types de bruits de bouche produits, en fonction de leur flux d'air, il apparaît une répartition par phase

[99] Il nous est possible de comparer les phases entre elles sans nous reporter à un taux de production par minute, car les différentes phases du scénario sont de durée globalement équivalente.

différente de la répartition globale dans chacun des cas (Tableau 11). 34% des bruits de bouche « égressifs bloqués » sont produits en phase 1. Au contraire, 30% des bruits de bouche « égressifs continus » sont produits en phase 3, avec dans chacun des cas, la production des autres bruits de bouche de même type bien répartie dans les trois autres phases.

En parallèle, la majorité des bruits de bouche produits dans un flux d'air ingressif et continu le sont au cours de la phase 2 (29%), alors que c'est aussi la phase où sont produits le moins de bruits de bouche « ingressifs bloqués ».

Tableau 11: Pourcentage de bruits de bouche par phase, selon leur flux d'air.

| Flux d'air | | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|---|
| egressif | bloqué | 34% | 21% | 23% | 22% | 100% |
| | gêné | 12% | 20% | 39% | 29% | 100% |
| | continu | 23% | 24% | 30% | 23% | 100% |
| ingressif | bloqué | 25% | 20% | 27% | 28% | 100% |
| | gêné | 10% | 23% | 40% | 27% | 100% |
| | continu | 25% | 29% | 22% | 24% | 100% |

Quant aux bruits de bouche produits avec un flux d'air gêné, ils sont principalement produits en phase 3, qu'ils soient ingressifs ou égressifs (39% et 40%, respectivement pour « égressif » et « ingressif » -Tableau 11-). De plus, leur nombre par phase suit dans les deux cas un même patron : augmentation de leur nombre de la phase 1 à la phase 2, et 2 à 3 où il atteint son maximum, puis légère diminution entre les phases 3 et 4.

Tableau 12: Nombre de bruits de bouche par phase selon leur flux d'air

| Flux d'air | | 1 | 2 | 3 | 4 | Total | |
|---|---|---|---|---|---|---|---|
| egressif | bloqué | 82 | 50 | 55 | 52 | 239 | 745 |
| | gêné | 8 | 14 | 27 | 20 | 69 | |
| | continu | 101 | 105 | 130 | 101 | 437 | |
| ingressif | bloqué | 217 | 175 | 231 | 238 | 861 | 1374 |
| | gêné | 5 | 12 | 21 | 14 | 52 | |
| | continu | 117 | 134 | 101 | 109 | 461 | |

Par ailleurs, nous avons vu dans la partie I.2. que les bruits de bouche sont plus ou moins fréquents selon leur flux d'air[100]. Pour chacune des phases les proportions des différents type de flux d'air sont globalement respectées. Par exemple dans toutes les phases, les bruits de bouche sont majoritairement produits dans un flux d'air ingressif,

[100] Dans le Tableau 12, le nombre total de bruits de bouche est de 2119 (et non 2358), car les 239 événements « mouillés » (« déglutit » et « interaction langue-lèvres ») dont le flux d'air est noté « non lieu » ne sont pas pris en compte.

et en particulier bloqué (ce qui correspond essentiellement aux relâchements d'articulateur et aux clicks).

En ce qui concerne les interjections (Tableau 13), leur répartition par phase selon leur type peut être très éloignée de la tendance globale du nombre d'interjections par phase décrite p.221. Cette répartition peut même montrer des tendances différentes entre un type d'interjection et sa variante (Tableau 13). Par exemple alors que 46% des « Comb » sont produites en phase 3, les « Comb_variantes » sont principalement produites en phase 1 (la moitié, contre seulement 13% des « Comb »)).

Plus localement, aucune variante de « Comb », ni de « CV » n'est produite dans la phase 2. Elles sont concentrées pour la moitié d'entre elles, respectivement dans les phases 1 et 4. De même, aucune variante de « VC » n'est produite au cours de la phase 3. Notons également que peu d'interjections « C » et « C_variante » sont produites en phase 4 (respectivement 13 et 9%), alors que globalement, 28% des interjections sont produites dans cette phase (21 à 50% des autres types d'interjections).

Tableau 14: Statistiques descriptives sur la production d'interjections par phase, selon leur type

| Phase | Moyenne des phases | Ecart-type | Coef. de variation |
|---|---|---|---|
| C | 5,75 | 1,89 | 0,33 |
| C_variante | 2,75 | 1,26 | 0,46 |
| Comb | 6 | 3,46 | 0,58 |
| Comb_variante | 1,5 | 1,29 | 0,86 |
| CV | 12 | 2,45 | 0,20 |
| CV_variante | 1,00 | 0,82 | 0,82 |
| V | 61,25 | 13,94 | 0,23 |
| V_variante | 17,75 | 3,86 | 0,22 |
| VC | 12,25 | 4,86 | 0,40 |
| VC_variante | 2,25 | 1,50 | 0,67 |
| Total | **122,5** | **21,30** | **0,17** |

Globalement, dans la distribution par phase des interjections, il existe une grande divergence pour les interjections les plus complexes articulatoirement (Tableau 14) : les interjections « Comb », les variantes des « Comb », « CV » et « VC » ont un coefficient de variation par phase compris entre 0,58 et 0,86 (contre 0,2 à 0,46 pour les autres types d'interjections).

Quant aux proportions des différents types d'interjections par phase, elles respectent globalement les tendances observées sur l'ensemble du scénario (voir Tableau 15).

Toutefois, le pourcentage d'interjections « Comb » est particulièrement important pendant la phase 3 (10% des événements vocaux pour la phase 3, contre 5% en moyenne sur tout le scénario) et celui des interjections « VC » est particulièrement faible en phase 1 (6% des événements dans cette phase, contre 10% sur l'ensemble du scénario).

Tableau 15: Pourcentage d'interjections par type,
selon la phase dans laquelle elles sont produites

| Phase | 1 | 2 | 3 | 4 | **Total** |
|---|---|---|---|---|---|
| C | 6% | 5% | 6% | 2% | **5%** |
| C_variante | 3% | 2% | 4% | 1% | **2%** |
| Comb | 3% | 3% | 10% | 4% | **5%** |
| Comb_variante | 3% | 0% | 1% | 1% | **1%** |
| CV | 9% | 8% | 11% | 11% | **10%** |
| CV_variante | 1% | 0% | 1% | 1% | **1%** |
| V | 51% | 52% | 44% | 51% | **50%** |
| V_variante | 15% | 15% | 13% | 15% | **14%** |
| VC | 6% | 12% | 10% | 11% | **10%** |
| VC_variante | 3% | 2% | 0% | 2% | **2%** |
| Total | 100% | 100% | 100% | 100% | **100%** |

## *II.2. Lien aux tâches récurrentes*

Lors de l'étiquetage, nous avons noté quelles sont les tâches récurrentes (que nous appellerons « tâches globales »), liées au scénario, dans lesquelles le sujet se trouve (Chapitre 3 IV.3.). Rappelons que leur valeur peut être « réponses », « sous-résultats » ou « warning ».

54% des événements vocaux sont produits pendant les tâches de type « réponses » : 1497 bruits de bouche et 55 interjections. Ainsi, 96% des événements vocaux de ce type de tâche globale sont des bruits de bouche. Le taux global moyen d'événements par minute est de 9,4 (Tableau 16).

Pendant ce type de tâche globale se trouvent 63% des bruits de bouche et 11% des interjections, soit en moyenne 9,1 bruits de bouche par minute, mais seulement 0,3 interjections.

Pendant les tâches de « sous-résultats », se trouvent 20% des bruits de bouche et 48% des interjections. En moyenne, 15,1 bruits de bouche, et 7,4 interjections sont produits par minute. D'autre part, 67% des événements vocaux de ce type de tâche globale sont des bruits de bouche.

Pendant les tâches de type « warning » se trouvent 16% des bruits de bouche et 41% des interjections. En moyenne, 13,6 bruits de bouche bouche, et 7,1 interjections sont

produits par minute. De plus, 66% des événements vocaux de ce type de tâche globale sont des bruits de bouche.

Il est intéressant de constater qu'il existe une différence importante entre les tâches globales de type « réponses » et celles de type « sous-résultats » et « warning », si nous considérons le rapport du nombre de bruits de bouche sur le nombre d'interjections : ce dernier est presque 15 fois plus important en « réponses » (avec un rapport de 27,2) qu'en « sous-résultats » (rapport de 2) ou en « warning » (1,9)

Tableau 16: Nombre (en haut) et taux moyen par minute (en bas) d'événements vocaux selon la tâche globale lors de laquelle ils sont produits et en fonction de leur catégorie

| Tâche globale | Bruits de bouche | Interjections | Total |
|---|---|---|---|
| Réponses | 1497 | 55 | **1552** |
| Sous-résultats | 478 | 236 | **714** |
| Warning | 383 | 199 | **582** |
| **Total** | **2358** | **490** | **2848** |

| Tâche globale | Bruits de bouche | Interjections | Total |
|---|---|---|---|
| Réponses | 9,1 | 0,3 | **9,4** |
| Sous-résultats | 15,1 | 7,4 | **22,5** |
| Warning | 13,6 | 7,1 | **20,7** |
| **Total** | **10,5** | **2,2** | **12,6** |

Globalement, la distinction entre tâches globales de type « sous-résultats » et de type « warning » n'apparaît pas pertinente à la vue de ces analyses. En effet, la différence de nombre d'occurrences existant entre ces deux types de tâches globales est faible. Toutefois, si nous ramenons à une fréquence, le nombre moyen d'occurrences par minute en fonctions de ces tâches globales (Tableau 17), il est possible de déterminer un comportement par phase plus spécifique pour chacune d'elle.

Les nombres moyens d'événements vocaux par minute les plus élevés concernent les tâches globales de type « sous-résultats », en particulier en phases 2 et 3 (avec respectivement 25,7 et 24,6 événements par minute). Il est à noter que ce sont les deux phases où la productions d'événements vocaux est la plus faible lors des tâches globales de type « warning » (avec respectivement 10,7 et 2,4 événements par minute en moyenne). Quant aux tâches globales de type « réponses », leur taux moyen d'événements vocaux par minute est relativement stable et faible : plus de deux fois moins important qu'en « sous-résultats », avec des taux de 8,3 à 10,8 événements par minute.

Les inductions émotionnelles provoquées par les tâches globales de type « sous-résultats » semblent jouer un rôle en augmentant globalement le taux d'événements vocaux produits par rapport aux tâches globales de type « réponses », qui ne

possèdent pas de composante émotionnelle induite. Toutefois, aucune corrélation ne peut être établie avec le type d'induction (c'est-à-dire se voulant positive *vs.* négative, selon le scénario).

Tableau 17: Taux moyen par minute (en bas) d'événements vocaux selon la tâche globale lors de laquelle ils sont produits et en fonction de la phase

| Tâche globale | Phase 1 | Phase 2 | Phase 3 | Phase 4 | Total |
|---|---|---|---|---|---|
| Réponses | 8,3 | 10,8 | 9,7 | 10,7 | **9,4** |
| Sous-résultats | 20,8 | 25,7 | 24,6 | 22,8 | **22,5** |
| Warning | 16,9 | 10,7 | 2,4 | 21,1 | **20,7** |
| **Total** | **10,7** | **15,2** | **13,6** | **14,5** | **12,6** |

Les tâches de type « warning » possèdent également une composante d'induction émotionnelle. Pourtant, le taux d'événements reste moins élevé qu'en tâche « sous-résultats », et surtout beaucoup plus variable selon la phase (de 2,4 à 21,1 événements par minute).

## *II.3. Lien à la nature de la tâche*

Au cours des tâches globales de type « réponses », les bruits de bouche sont produits principalement avant de répondre (26%), pendant une réponse (24%) ou pendant la lecture des consignes (24%). Aucune tendance ne se dégage pour les interjections dans ce type de tâche globale.

À l'inverse, lors des « sous-résultats » et « warning », respectivement 48% et 55% des bruits de bouche, mais surtout 89% et 96% des interjections, sont produits pendant les commentaires, et non pendant la lecture.

Il est à souligner qu'il existe, de par notre scénario, deux types de tâche de lecture différentes. Cette différence dépend de la tâche globale dans laquelle le sujet se trouve : lors des tâches globales de type « réponses », il s'agit d'une lecture de consignes ou d'indications, donc d'une tâche de compréhension ; lors des tâches globales « sous-résultats » et « warning », il s'agit au contraire d'une lecture de résultats. Elle place les sujets devant une évaluation (fausse) de leur compétence, dont le caractère négatif ou positif est le facteur d'induction : un changement d'états affectifs (attendu et vérifié) découle donc de la lecture. Cette composante émotionnelle de la tâche pourrait expliquer en partie les différences dans le nombre et le type d'événements vocaux produits, entre ces deux types de lecture.

## *II.4. Lien à la prise de parole*

Tableau 18: Pourcentage d'événements vocaux selon leur position par rapport au(x) prise(s) de parole et en fonction de leur catégorie

| Lien à la parole | Bruits de bouche | Interjections | Total |
|---|---|---|---|
| **0** | 36% | 7% | **31%** |
| **1** | 13% | 62% | **21%** |
| **~** | 15% | 9% | **14%** |
| **Avant** | 28% | 19% | **26%** |
| **Après** | 9% | 2% | **8%** |
| **Total** | 100% | 100% | **100%** |

Une fois de plus, le comportement des bruits de bouche se distingue de celui des interjections quant à leur position par rapport aux prises de parole du sujet (voir Chapitre 5 III.1.). 36% des bruits de bouche sont produits en dehors de toute prise de parole, et 28% avant une prise de parole (Tableau 18). À l'inverse, 62% des interjections sont produites pendant une prise de parole (contre 13% des bruits de bouche).

Plus précisément, la position temporelle des bruits de bouche par rapport aux prises de parole semble dépendre de leur type de flux d'air (Tableau 19[101]).

Tableau 19: Nombre de bruits de bouche par type de flux d'air en fonction de leur position aux prises de parole

| | Type de flux d'air | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Relation à la prise de parole** | Egressif | | | Ingressif | | | | | |
| | bloqué | gêné | continu | bloqué | gêné | continu | Déglutit | Interaction langue-lèvres | **Total** |
| 0 | 109 | 25 | 113 | 308 | 13 | 124 | 96 | 52 | **840** |
| 1 | 16 | 25 | 227 | 28 | 1 | 4 | 1 | 1 | **303** |
| ~ | 27 | 6 | 24 | 87 | 27 | 164 | 13 | 0 | **348** |
| Before | 25 | 8 | 32 | 416 | 10 | 123 | 27 | 13 | **654** |
| After | 62 | 5 | 41 | 22 | 1 | 46 | 30 | 6 | **213** |
| **Total** | **239** | **69** | **437** | **861** | **52** | **461** | **167** | **72** | **2358** |
| | 745 | | | 1374 | | | | | |

Nous observons, parmi les bruits de bouche produits dans un fux d'air égressif, que 52% des bruits de bouche « continus », mais 7% des bruits de bouche « bloqués », apparaissent pendant une production de parole. 46% de ces bruits de bouche « bloqués » sont en effet produits en dehors de toute prise de parole.

Quant aux bruits de bouche produits dans un flux d'air égressif gêné, 72% d'entre eux sont produits soit pendant, soit en dehors des productions de parole. Les relations à la

[101] Une précision quant aux comptages des événements dans ces tableaux : les catégories « déglutit » et « interaction langue-lèvres mouillées » sont comptés seulement lorsqu'ils apparaissent seuls, c'est-à-dire lorsque leur paramètre « type de flux d'air » est noté « non lieu » (voir Chapitre 5 III.2.).

prise de parole notées « ~ » (entre deux productions), « Before » (avant) et « After » (après) ne concernent, à elles trois, que 31% des bruits de bouche produits dans un flux d'air égressif.

En parallèle, seulement 2% des bruits de bouche produits dans un flux d'air ingressif (quel qu'il soit) sont produits pendant un prise de parole. La raison est certainement que le français est une langue dont les phonèmes sont égressifs, et que la production d'un bruit de bouche ingressif pendant une prise de parole nécessite un changement de type de flux d'air, et donc un contrôle plus important.

Les bruits de bouche ingressifs sont ainsi principalement produits :

- lorsqu'ils sont « bloqués », avant (pour 48%) ou en dehors (pour 36%) d'une prise de parole ;

- 36% des « continus » entre deux prises de parole (noté « ~ »), de même que 52% des « gênés ».

Quant aux bruits de bouche « de déglutition » et « d'interaction langue-lèvres », respectivement 57% et 72% apparaissent en dehors de toute prise de parole.

Si nous considérons indépendamment chaque type de relation aux productions de parole, nous remarquons que :

- 52% des bruits de bouche produits pendant une prise de parole ont leur flux d'air égressif et continu ;
- 47% des bruits de bouche produits entre deux prises de parole ont leur flux d'air ingressif et continu ;
- 82% des bruits de bouche produits avant une prise de parole ont leur flux d'air ingressif (et 48% des « ingressifs bloqués » à eux seuls) ;
- 50% des bruits de bouche produits après une prise de parole ont leur flux d'air égressif (quel qu'il soit).

## *II.5. Lien entre les événements vocaux*

Le Tableau 20 montre quels sont les types de bruits de bouche qui sont les constituants d'événements « doubles ». Précisons que lorsqu'un « élément qui suit » est noté, ce dernier est le deuxième élément d'un événement, dont les deux constituants sont produits dans un même flux d'air (voir les critères Chapitre 5 III.2.). Ce tableau ne prend donc pas en compte la succession d'éléments indépendants.

Les différents couples d'événements possibles ne sont pas tous fréquents, ni même produits. Concernant les tendances, 103 sur 163 (soit 63%) événements « doubles »

commencent par une déglutition ou une inspiration. Quant aux éléments qui suivent, il s'agit principalement de relâchements d'articulateur (26%) et d'inspirations (21%).

Tableau 20: Éléments constituants d'événements « doubles »

| | Elément qui suit | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Type de bruits de bouche | déglutit | click | expiration | friction | inspiration | langue-lèvres mouillées | occlusion | occlusion ingressive _inspiration | relâche sa respiration | relâchement articulateur | **Total** |
| déglutit | | 8 | 3 | | 4 | | 5 | 4 | 18 | 10 | **52** |
| click | | | | | 11 | | | | | | **11** |
| expiration | 1 | | | | | | | | | 3 | **4** |
| inspiration | | 3 | | 1 | | 5 | 4 | 12 | | 26 | **51** |
| langue-lèvres mouillées | | 2 | 1 | | 9 | | | 2 | | | **14** |
| occlusion | | | 2 | | | | 3 | 1 | | 1 | **7** |
| occlusion ingressive_insp | | | | | | | 1 | 1 | | | **2** |
| relâche sa respiration | | | | | | 1 | | | | 1 | **2** |
| relâchement articulateur | | | 1 | | 11 | | 4 | 2 | | 2 | **20** |
| **Total** | **1** | **13** | **7** | **1** | **35** | **6** | **17** | **22** | **18** | **43** | **163** |

Par ailleurs, certains types de bruits de bouche ne sont jamais produits au sein d'un couple : « expiration brutale », « gémissement », « raclement de gorge » et « friction » en première position.

Les couples les plus fréquents sont « inspiration + relâchement articulateur » (16%), « déglutit + relâche sa respiration » (11%), et « click + inspiration », « relâchement articulateur + inspiration » et « inspiration + occlusion ingressive_inspiration » (7% chaque).

Il reste à déterminer ce qui influe sur ces fréquences, c'est-à-dire à quels types de contraintes ces fréquences sont liés (physiologiques ? situationnelles ?).

Les différentes analyses de cette deuxième partie du chapitre permettent de justifier la distinction que nous avons faite entre bruits de bouche et interjections, à travers leurs différences de comportements en fonction des paramètres situationnels découlant de la nature de nos données. De la même manière, elles montrent la pertinence de nos distinctions des bruits de bouche par type de flux d'air, et des interjections par type acoustique et variantes (c'est-à-dire en termes de présence de qualité de voix).

# III. Le rôle des motifs temporels dans la caractérisation comportementale des personnes

Nous avons étudié en détail l'organisation des événements vocaux de six sujets du corpus, essentiellement en dehors de ses prises de parole. L'objectif est de se faire une première idée sur les facteurs susceptibles de faire varier le comportement vocal de chacun des sujets (avec ses particularités), et sur sa perception par autrui.

Pour cela, nous avons cherché l'existence de motifs temporels dans nos données. Ces motifs sont étudiés en tant que fréquence et régularité de comportements, en fonction des mêmes paramètres situationnels que ceux étudiés dans la partie II. de ce chapitre. Nous avons étudié à la fois :

- les variations intra-sujet, c'est-à-dire le comportement du sujet face à une situation particulière ;
- les variations inter-sujets, c'est-à-dire dans quelle mesure la personnalité de chacun des sujets va se traduire par un comportement expressif différent face à une même situation.

En effet, comme nous l'avons introduit Chapitre 5 III.3., nous adaptons notre perception des comportements au sujet qui les produit. Lors de l'étiquetage, et probablement, par généralisation à la vie de tous les jours, lorsque nous rencontrons de nouvelles personnes, notre perception va s'adapter au sujet / à la personne : par comparaison aux autres (sujets) ; mais surtout par comparaison au comportement expressif global propre au sujet / à la personne dans la situation donnée.

Par exemple si un sujet est très silencieux et bouge très peu, la perception, même faible, des événements non- et pré-verbaux, va certainement être plus fine car elle va nous apparaître, en tant qu'humain communicant, comme plus informationnelle et pertinente au niveau expressif, que les mêmes éléments peu perceptibles chez une personne naïvement décrite comme expressive ou extravertie.

En quelque sorte, nous adaptons un « seuil de pertinence » à chaque personne, qui nous permet de relativiser la quantité absolue par rapport à la variation.

Il y a donc en jeu un apprentissage sur ce que nous percevons du sujet, à partir de son comportement expressif et de la personnalité que nous lui attribuons naïvement dans la situation donnée. Cette considération nous permet de confirmer l'intérêt de ne pas chercher des prototypes expressifs d'événements vocaux, mais plutôt chercher à pointer puis modéliser la variation (*cf.* Chapitre 2 II.3.).

## *III.1. Effets inter-sujets et analyse globale*

Les six sujets ont utilisé la modalité vocale de manière différente (Figure 44). Aucune corrélation n'apparaît entre le nombre global d'événements vocaux et la réaction des sujets aux inductions de l'expérience, c'est-à-dire aux phases du scénario (Figure 44). Toutefois leur taux relatif par sujet pendant les trois dernières phases de l'expérience laisse apparaître deux groupes de sujets fonction de leur comportement : F_T, M_N, M_J, qui se sont réellement inquiétés sur la dégradation de leurs capacités ; et F_S, F_M, M_R, qui ont plutôt remis en question le logiciel d'apprentissage (personne n'a eu de doute par rapport au magicien d'Oz). La phase 1, étant une phase d'apprentissage de la tâche, sans induction émotionnelle recherchée, sépare quant à elle les stratégies par rapport à la tâche, de F_T, F_S, M_ R, M_N d'un côté *vs.* F_M, M_ J de l'autre.

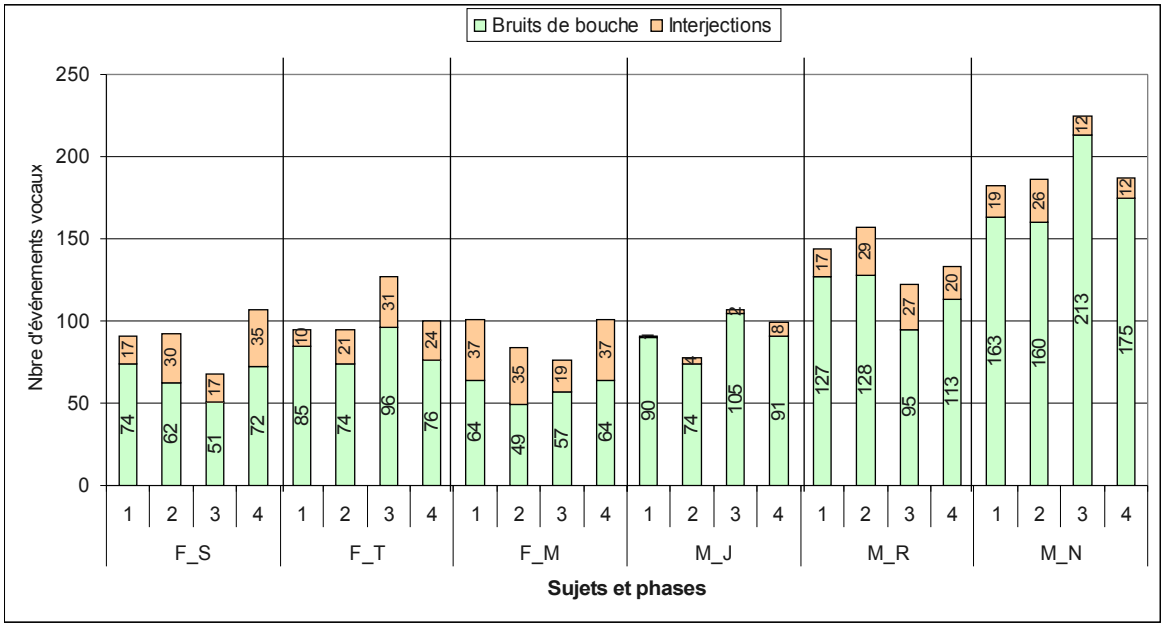



Figure 44: Comparaison inter-sujets du nombre d'événements vocaux (bruits de bouche et interjections) pendant les différentes phases du scénario.

Ces groupes ne peuvent être reliés à l'âge, au sexe, au niveau d'éducation, ni la familiarité du sujet avec les IHM. Un test préliminaire sur la personnalité perçue des sujets (en terme d'exubérance) par des juges naïfs, ne laisse apparaître aucune relation entre éléments triviaux d'intro-/ extraversion, mais cela reste à étudier.

Le nombre d'événements vocaux des hommes est globalement plus élevé que pour les femmes, mais le ratio interjections/autres événements vocaux est plus élevé pour les femmes (Figure 45).

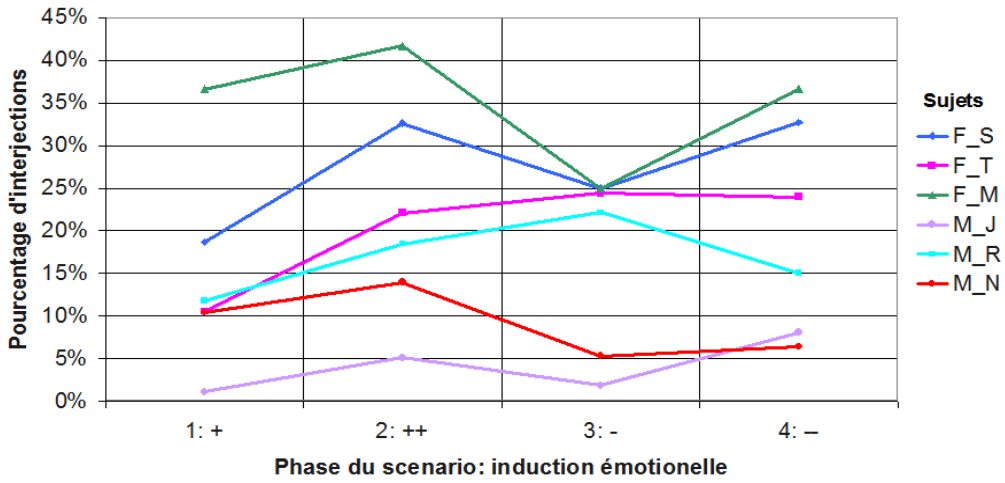



Figure 45: Comparaison inter-sujets du pourcentage d'interjections sur le nombre total d'événements vocaux, selon la phase (les signes + et – sont un rappel du renforcement -positif ou négatif- associé à la phase).

Cela peut être relié à la quantité de parole (commentaires libres) qui est également plus élevée chez les trois femmes que chez les trois hommes (*cf.* un tableau récapitulatif des données temporelles pour chacun des sujets en Annexe 10). Ces femmes parlent plus longtemps et utilisent plus d'interjections pré-lexicales par rapport aux autres bruits de bouche. Globalement, elles produisent aussi moins d'items non-lexicaux (Figure 46). Les hommes parlent moins et semblent favoriser la production d'items non phonémiques et non phonétiques.

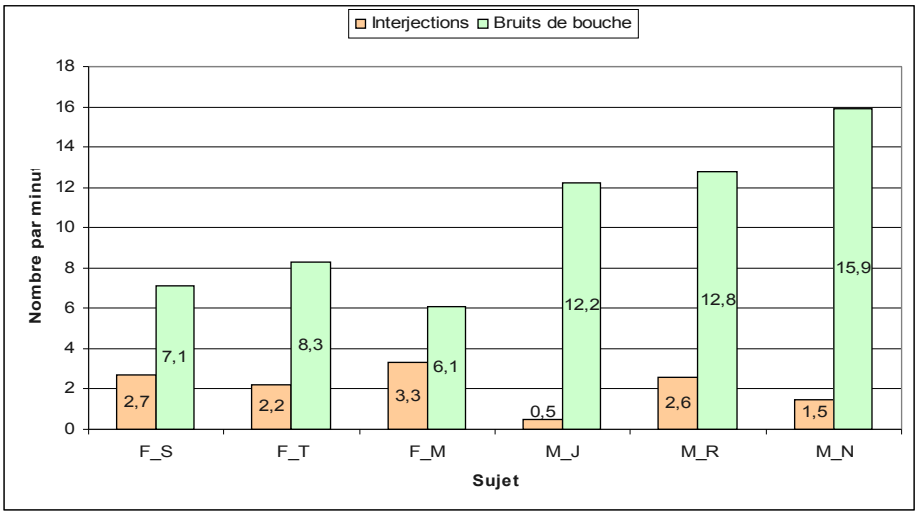



Figure 46: Comparaison inter-sujets du nombre moyen de bruits du bouche et d'interjections par minute.

Bien sûr, cette distinction homme/femme ne peut être généralisée avec six sujets (au moins un sujet masculin du corpus semble avoir les mêmes caractéristiques comportementales que les trois femmes), mais cette « tendance » au langage, avec peu de bruits de bouche et une proportion élevée d'interjections pourrait peut-être être évaluée en tant qu'élément d'une stratégie communicative particulière.

## *III.2. Nature des interjections*

Les répartitions temporelles des différents types d'interjections montrent de grandes divergences entre elles. Les grandes tendances montrent toutefois :

- que les interjections sont principalement des phonèmes vocaliques : entre 52 et 73% du nombre total d'interjections selon le sujet. Plus précisément, l'interjection la plus fréquente est l'interjection vocalique [ø:] (« euh ») pour tous les sujets et que soient incluses ou non les variantes d'interjections en terme de qualité de voix ;
- que la répartition par type des interjections consonantiques et celles plus complexes diffèrent selon les sujets et selon les phases du scénario : par exemple le second type d'interjections le plus fréquent est « VC » pour F_S et F_M (16%), « CV » pour M_N (25%), et « C » pour M_R (16%) ;
- que l'utilisation de variantes de qualité de voix est également dépendant des sujets, que ce soit en termes de fréquence ou de type utilisé (Tableau 21).

Tableau 21: Comparaison inter-sujets du nombre d'interjections selon leur nature, en fonction de la phase du scénario.

| Type d'interjections | C | | | | | | V | | | | | | CV | | | | | | VC | | | | | | Comb | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sujet | F_S | F_T | F_M | M_J | M_R | M_N | F_S | F_T | F_M | M_J | M_R | M_N | F_S | F_T | F_M | M_J | M_R | M_N | F_S | F_T | F_M | M_J | M_R | M_N | F_S | F_T | F_M | M_J | M_R | M_N | |
| Phase 1 | 1 | 1 | 2 | 0 | 2 | 0 | 9 | 8 | 17 | 1 | 3 | 9 | 3 | 0 | 4 | 0 | 1 | 1 | 0 | 0 | 4 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 69 |
| Phase 2 | 0 | 2 | 1 | 0 | 4 | 0 | 18 | 14 | 15 | 1 | 10 | 14 | 2 | 1 | 1 | 0 | 2 | 5 | 7 | 0 | 8 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 3 | 110 |
| Phase 3 | 0 | 6 | 0 | 0 | 1 | 0 | 11 | 11 | 11 | 0 | 9 | 3 | 2 | 1 | 1 | 0 | 2 | 6 | 3 | 1 | 4 | 1 | 1 | 1 | 1 | 2 | 1 | 0 | 6 | 1 | 86 |
| Phase 4 | 1 | 0 | 1 | 0 | 1 | 0 | 24 | 12 | 18 | 2 | 6 | 2 | 2 | 0 | 4 | 0 | 4 | 3 | 4 | 4 | 3 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 3 | 0 | 98 |
| **Total** | **2** | **9** | **4** | **0** | **8** | **0** | **62** | **43** | **61** | **4** | **28** | **28** | **9** | **2** | **10** | **0** | **9** | **15** | **14** | **5** | **19** | **1** | **3** | **4** | **1** | **3** | **3** | **0** | **10** | **4** | **361** |
| **Total par type** | **23** | | | | | | **226** | | | | | | **45** | | | | | | **46** | | | | | | **21** | | | | | | |

| Type d'interjections | C-variante | | | | | | V-variante | | | | | | CV-variante | | | | | | VC-variante | | | | | | Comb-variante | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sujet | F_S | F_T | F_M | M_J | M_R | M_N | F_S | F_T | F_M | M_J | M_R | M_N | F_S | F_T | F_M | M_J | M_R | M_N | F_S | F_T | F_M | M_J | M_R | M_N | F_S | F_T | F_M | M_J | M_R | M_N | |
| Phase 1 | 0 | 0 | 1 | 0 | 2 | 0 | 3 | 1 | 8 | 0 | 3 | 5 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 2 | 2 | 32 |
| Phase 2 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 3 | 8 | 2 | 10 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 35 |
| Phase 3 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 10 | 2 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 22 |
| Phase 4 | 0 | 0 | 1 | 0 | 0 | 0 | 3 | 6 | 6 | 4 | 5 | 4 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 2 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 40 |
| **Total** | **1** | **0** | **3** | **0** | **7** | **0** | **7** | **20** | **24** | **7** | **20** | **12** | **0** | **1** | **1** | **1** | **2** | **2** | **2** | **3** | **1** | **1** | **3** | **2** | **1** | **0** | **2** | **1** | **3** | **2** | **129** |
| **Total par type** | **11** | | | | | | **90** | | | | | | **7** | | | | | | **12** | | | | | | **9** | | | | | | |

## *III.3. Nature des bruits de bouche*

Les analyses indiquent qu'il existe des tendances globales, pour la majorité des sujets, en termes de répartition par type, dans leur production de bruits de bouche (Tableau 22[102]). C'est notamment le cas de la répartition des bruits de bouche en termes de flux d'air ingressif ou égressif d'une part, et bloqué, gêné ou continu (pour une description des critères, *cf.* Chapitre 5 III.2.). Ainsi, le type de bruits de bouche le plus fréquent est le type « bruits de bouche bloqué » produit dans un fux d'air ingressif (entre 35 et 51% du nombre total de bruits de bouche), pour tous les sujets à l'exception de M_J. Ce dernier produit le plus fréquemment des bruits de bouche égressifs continus (35% de ses événements vocaux).

[102] Comme pour le Tableau 19, les catégories « déglutit » et « interaction langue-lèvres mouillées » sont comptés seulement lorsqu'ils apparaissent seuls, c'est-à-dire lorsque leur paramètre « type de flux d'air » est noté « non lieu » (voir Chapitre 5 III.2.).

Tableau 22: Comparaison inter-sujets du nombre de bruits de bouche selon leur nature, et en fonction de leur position temporelle par rapport à la production de parole :
partie haute, les bruits de bouche produits dans un flux d'air égressif ou ingressif ;
partie basse, les bruits de bouche produits sans un flux d'air et rappel du nombre total de bruits de bouche selon leur position par rapport aux prises de parole.
Chaque partie est ensuite divisée en sujets féminins en haut, et masculins en bas.

| **Sujet Féminin** | Egressif | | | | | | | | | Ingressif | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bloqué | | | gêné | | | continu | | | bloqué | | | gêné | | | continu | | |
| Sujet F_ | **T** | **S** | **M** | **T** | **S** | **M** | **T** | **S** | **M** | **T** | **S** | **M** | **T** | **S** | **M** | **T** | **S** | **M** |
| 0 | 17 | 8 | 9 | 2 | 3 | 1 | 21 | 6 | 3 | 40 | 23 | 19 | 3 | 0 | 3 | 9 | 3 | 5 |
| 1 | 2 | 0 | 1 | 12 | 1 | 1 | 44 | 15 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| ~ | 4 | 13 | 2 | 0 | 0 | 1 | 3 | 7 | 1 | 19 | 12 | 30 | 0 | 13 | 0 | 40 | 36 | 21 |
| Before | 7 | 0 | 5 | 1 | 0 | 3 | 2 | 1 | 4 | 56 | 49 | 64 | 0 | 1 | 1 | 11 | 7 | 4 |
| After | 3 | 2 | 6 | 2 | 2 | 0 | 5 | 1 | 2 | 1 | 3 | 5 | 0 | 0 | 0 | 4 | 9 | 3 |
| **Total** | **33** | **23** | **23** | **17** | **6** | **6** | **75** | **30** | **11** | **116** | **87** | **119** | **3** | **14** | **4** | **64** | **56** | **34** |
| | 79 | | | 29 | | | 116 | | | 322 | | | 21 | | | 154 | | |
| | **224** | | | | | | | | | **497** | | | | | | | | |

| **Sujet Masculin** | Egressif | | | | | | | | | Ingressif | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bloqué | | | gêné | | | continu | | | bloqué | | | gêné | | | continu | | |
| Sujet M_ | **R** | **J** | **N** | **R** | **J** | **N** | **R** | **J** | **N** | **R** | **J** | **N** | **R** | **J** | **N** | **R** | **J** | **N** |
| 0 | 18 | 13 | 44 | 2 | 5 | 12 | 39 | 17 | 27 | 83 | 41 | 102 | 2 | 4 | 1 | 22 | 15 | 70 |
| 1 | 0 | 4 | 9 | 0 | 2 | 9 | 5 | 99 | 63 | 1 | 3 | 23 | 0 | 0 | 1 | 1 | 0 | 1 |
| ~ | 2 | 0 | 6 | 4 | 1 | 0 | 1 | 2 | 10 | 9 | 6 | 11 | 10 | 3 | 1 | 37 | 6 | 24 |
| Before | 2 | 5 | 6 | 1 | 2 | 1 | 17 | 3 | 5 | 87 | 46 | 114 | 1 | 5 | 2 | 33 | 26 | 42 |
| After | 6 | 12 | 33 | 1 | 0 | 0 | 11 | 4 | 18 | 3 | 2 | 8 | 0 | 1 | 0 | 8 | 15 | 7 |
| **Total** | **28** | **34** | **98** | **8** | **10** | **22** | **73** | **125** | **123** | **183** | **98** | **258** | **13** | **13** | **5** | **101** | **62** | **144** |
| | 160 | | | 40 | | | 321 | | | 539 | | | 31 | | | 307 | | |
| | **521** | | | | | | | | | **877** | | | | | | | | |

| **Sujet Féminin** | Sans flux d'air | | | | | |
|---|---|---|---|---|---|---|
| | Déglutit | | | Interaction langue-lèvres | | |
| Sujet F_ | **T** | **S** | **M** | **T** | **S** | **M** |
| 0 | 11 | 16 | 9 | 8 | 9 | 7 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| ~ | 1 | 0 | 4 | 0 | 0 | 0 |
| Before | 1 | 4 | 8 | 1 | 2 | 0 |
| After | 0 | 12 | 5 | 1 | 0 | 4 |
| **Total** | **13** | **32** | **26** | **10** | **11** | **11** |
| | 71 | | | 32 | | |

| **Sujets Féminins** | **Total: Tout type de flux d'air** |
|---|---|
| 0 | 175 |
| 1 | 80 |
| ~ | 202 |
| Before | 216 |
| After | 48 |
| **Total** | **721** |

| **Sujet Masculin** | Sans flux d'air | | | | | |
|---|---|---|---|---|---|---|
| | Déglutit | | | Interaction langue-lèvres | | |
| Sujet M_ | **R** | **J** | **N** | **R** | **J** | **N** |
| 0 | 21 | 5 | 34 | 18 | 6 | 4 |
| 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| ~ | 3 | 0 | 5 | 0 | 0 | 0 |
| Before | 5 | 0 | 9 | 2 | 3 | 5 |
| After | 7 | 3 | 3 | 0 | 1 | 0 |
| **Total** | **36** | **8** | **52** | **21** | **10** | **9** |

| **Sujets Masculins** | **Total: Tout type de flux d'air** |
|---|---|
| 0 | 517 |
| 1 | 221 |
| ~ | 133 |
| Before | 398 |
| After | 129 |
| **Total** | **1398** |

Pour tous les sujets, les bruits de bouche ingressifs sont principalement produits avec un flux d'air bloqué (de 55 à 76% des bruits de bouche ingressifs selon les sujets) alors que les bruits de bouche égressifs sont plutôt produits avec un flux d'air continu (de 51 à 74% des bruits de bouche égressifs). Seule F_M produits majoritairement ses bruits de bouche égressifs avec un flux d'air bloqué (58% d'entre eux).

Les bruits de bouche sont peu produits avec un flux d'air « gêné ». Lorsqu'ils le sont, leur flux d'air est principalement égressif et ils représentent entre 10 et 15% des bruits de bouche égressifs chez les femmes, et de 6 à 9% chez les hommes. Toutefois, leur faible fréquence empêche de dégager des tendances plus précises.

Tableau 23: Comparaison inter-sujets du pourcentage de bruits de bouche selon leur type de flux d'air, et en fonction de leur position temporelle par rapport à la production de parole (sujets féminins en haut, et masculins en bas)

| **Sujets Féminins** | Egressif | | | | | | | | | Ingressif | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bloquée | | | gênée | | | continue | | | bloquée | | | gênée | | | continue | | |
| F_ | T | S | M | T | S | M | T | S | M | T | S | M | T | S | M | T | S | M |
| 0 | 52% | 35% | 39% | 12% | 50% | 17% | 28% | 20% | 27% | 34% | 26% | 16% | 100% | 0% | 75% | 14% | 5% | 15% |
| 1 | 6% | 0% | 4% | 71% | 17% | 17% | 59% | 50% | 9% | 0% | 0% | 1% | 0% | 0% | 0% | 0% | 2% | 3% |
| ~ | 12% | 57% | 9% | 0% | 0% | 17% | 4% | 23% | 9% | 16% | 14% | 25% | 0% | 93% | 0% | 63% | 64% | 62% |
| Before | 21% | 0% | 22% | 6% | 0% | 50% | 3% | 3% | 36% | 48% | 56% | 54% | 0% | 7% | 25% | 17% | 13% | 12% |
| After | 9% | 9% | 26% | 12% | 33% | 0% | 7% | 3% | 18% | 1% | 3% | 4% | 0% | 0% | 0% | 6% | 16% | 9% |
| **Total** | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

| **Sujets Masculins** | Egressif | | | | | | | | | Ingressif | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bloquée | | | gênée | | | continue | | | bloquée | | | gênée | | | continue | | |
| M_ | R | J | N | R | J | N | R | J | N | R | J | N | R | J | N | R | J | N |
| 0 | 64% | 38% | 45% | 25% | 50% | 55% | 53% | 14% | 22% | 45% | 42% | 40% | 15% | 31% | 20% | 22% | 24% | 49% |
| 1 | 0% | 12% | 9% | 0% | 20% | 41% | 7% | 79% | 51% | 1% | 3% | 9% | 0% | 0% | 20% | 1% | 0% | 1% |
| ~ | 7% | 0% | 6% | 50% | 10% | 0% | 1% | 2% | 8% | 5% | 6% | 4% | 77% | 23% | 20% | 37% | 10% | 17% |
| Before | 7% | 15% | 6% | 13% | 20% | 5% | 23% | 2% | 4% | 48% | 47% | 44% | 8% | 38% | 40% | 33% | 42% | 29% |
| After | 21% | 35% | 34% | 13% | 0% | 0% | 15% | 3% | 15% | 2% | 2% | 3% | 0% | 8% | 0% | 8% | 24% | 5% |
| **Total** | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

Nous avons vu dans la partie II.4. que la répartition globale des bruits de bouche en fonction de leur position temporelle par rapport aux prises de parole apparaît pertinente, puisqu'il est possible de dégager des tendances globales propres à certains types de bruits de bouche. En même temps que les tendances globales, les Tableaux 22 et 23 montrent également que cette même répartition est variable selon les sujets.

Les bruits de bouche « égressifs bloqués » sont surtout produits :

- en dehors de toute production de parole (« 0 ») par F_T, M_R et M_N (45 à 64% des bruits de bouche de ce type) ;
- répartis entre « 0 » et « After » pour M_J, et entre « 0 » et « After » et « Before » dans une moindre mesure pour F_M ;
- entre deux prises de parole « ~ » pour F_S (57% d'entre eux).

Les bruits de bouche « égressifs continus » sont essentiellement produits pendant les productions de parole (« 1 »), sauf pour M_R (53% d'entre eux sont produits en dehors

de toute production de parole (« 0 »), et pour F_M (36% avant les productions de parole, et 27% en dehors.

Les bruits de bouche « ingressifs bloqués » sont principalement produits avant les productions de parole (« Before ») par les trois femmes (de 48 à 56% d'entre eux), mais essentiellement répartis entre avant (« Before », pour 44 à 48% d'entre eux), et en dehors (« 0 », pour 40 à 45% d'entre eux) des productions de parole pour les trois hommes.

Les bruits de bouche « ingressifs continus » sont surtout produits :

- entre deux prises de parole « ~ » pour les trois femmes (62% à 64% d'entre eux) ;
- en dehors de toute production de parole (« 0 ») pour M_N (49% d'entre eux) ;
- avant une prise de parole (« Before ») pour M_J (42% d'entre eux) ;
- répartis entre « Before » et « ~ » pour M_R (respectivement 33 et 37%).

Le faible nombre de bruits de bouche produits avec un flux d'air gêné ne nous permet pas d'interpréter leur répartition. Notons seulement :

- que 17% des égressifs sont produits par F_T pendant ses productions de parole, et 17% et 13% par M_N, respectivement en dehors et pendant ses productions de parole. Chacun des autres cas (de couple position / sujet) ne dépasse pas 7% ;
- que 25% et 19% des ingressifs sont produits entre deux productions de parole, respectivement par F_S et M_R. Chacun des autres cas (de couple position / sujet) ne dépasse pas 10%.

Quant aux bruits de bouche sans flux d'air particulier (« déglutit » et « interaction langue-lèvres »), leur répartition est relativement stable entre les sujets, et diffère peu de leur tendance globale (*cf.* partie II.4.).

## *III.4. Perspectives technologiques et patrons de comportements*

Nous avons mis en évidence dans cette partie que les événements vocaux, et plus précisément leur différents types, étaient produits de manière différente selon les sujets. Les analyses confirment l'influence des paramètres situationnels et du profil psychologique des sujets sur le comportement vocal. Elles suggèrent ainsi l'existence de patrons comportementaux au niveau des expressions vocales, qui permettraient de caractériser les personnes, la situation étant donnée.

Par exemple si avant chacune de ses réponses, un sujet produit un bruit de bouche particulier, ce comportement sera peu pertinent pour interpréter les états de ce sujet dans ce contexte précis. Par contre, il sera une caractéristique propre au sujet, à sa personnalité, et à son humeur. Par ailleurs, c'est par ce type de comportement que, dans notre quotidienne, nous savons dans quel état d'esprit se trouve un ami à

l'instant T où nous le rencontrons. Ces comportements ont une valeur identitaire, dont nous mémorisons une représentation pour chacune de nos connaissances.

Au niveau des applications technologiques, un intérêt de cette observation est qu'il pourrait donc être possible de caractériser un locuteur, en termes de profil psychologique ou d'humeur, en étudiant ses patrons de comportement vocal.

En parallèle, dans le domaine des ACAs, un des objectifs actuels est que l'agent ait un comportement, en particulier d'écoute, différent en fonction de « sa personnalité » (*e.g.* de Sevin, Hyniewska, & Pelachaud, 2010), pour l'influence de la personnalité sur la production de *backchannels*). Il est par conséquent nécessaire de savoir quelles sont les personnalités les mieux perçues (*e.g.* celles dont les états exprimés sont les mieux reconnus) et qu'il est ainsi préférable de modéliser. Il serait donc intéressant de tester la perception de différents patrons comportementaux, puis de mettre en place des tests utilisateurs.

Nos observations ouvrent également une perspective de recherche pour les technologies de reconnaissance du locuteur. Les systèmes extraient le plus souvent un grand nombre de paramètres acoustiques des données sources et des potentiels locuteurs visés (*cf.* Fredouille, 2002 ; Fauve, Matrouf, Scheffer, Bonastre, & Mason, 2007) ; et les campagnes d'évaluation NIST[103]). Ils tentent ensuite d'apparier les informations au moyen de modèles probabilistes. Les recherches actuelles dans ce domaine portent actuellement, entre autres, sur le problème de la variabilité due au locuteur (*e.g.* ses émotions, sa fatigue, son stress.

De plus, Böhm & Shattuck-Hufnagel (2009) ont montré récemment qu'il existe une différence inter-locuteurs au niveau des types d'expiration produits en fin de production de parole : ce qu'ils appellent « speaker's habitual Utterance Final Phonation type »[104] ferait ainsi partie des caractéristiques de comportement vocal des locuteurs. Nous supposons quant à nous, que modéliser la manière dont un locuteur utilise les bruits de bouche et interjections dans une situation donnée pourrait être un paramètre utile à représenter dans un modèle de locuteur.

Cependant, avant de pouvoir tester cette hypothèse, il sera nécessaire au préalable d'être capable de détecter automatiquement ces événements, ce qui n'apparaît pas être une tâche aisée, à la vue de leur variété et de leur complexité acoustique (*cf.* Chapitre 7).

[103] http://www.itl.nist.gov/iad/mig/tests/mt/ site consulté pour la dernière fois le 28/02/20011

[104] littéralement « habitude du locuteur en termes de type de phonation en fin d'occurrence »

# IV. Perspectives autour des motifs temporels

## *IV.1. La notion de « niveaux temporels » et sa problématique*

De chaque type de transcription / de représentation découle les analyses postérieures (Mondada, 2008). Ainsi, la représentation des transcriptions « en ligne » est largement utilisée par les interactionnistes, et y compris ceux qui travaille sur la multimodalité. Son avantage est de mettre en évidence l'aspect séquentiel de l'interaction (Ochs, 1979). En effet, dans le domaine de l'analyse conversationnelle, le flux audio-visuel de l'interaction est considéré comme une alternance de tour de parole au cours de laquelle certains éléments non-verbaux « pertinents » pour l'interaction vont être produits. Dans cette mesure, c'est également comme telle que l'interaction va être transcrite (c'est-à-dire une alternance de tours de parole séparés par des pauses, sur laquelle seuls certains éléments non-verbaux vont être notés.

Ainsi, il s'agit dans une tâche de transcription ou d'étiquetage audio-visuel « en ligne », de se poser la question de l'étalon temporel : le « temps de référence » est-il donné par la parole ou par les gestes ?

Nous considérons quant à nous l'interaction communicative dans sa multimodalité. C'est pourquoi notre étiquetage a été effectué dans un éditeur d'annotations utilisant un système de champs. Cela nous permet de noter en parallèle les différentes modalités et les différents éléments / paramètres liés au scénario, de manière précise au niveau temporel, cela même lorsque les éléments étaient simultanés.

Ces réflexions font référence à une idée sous-jacente : la considération de différents niveaux temporels simultanés. Intuitivement, il serait possible de distinguer au moins trois niveaux temporels :

- le temps communicatif, temps de la parole rythmé par les réponses du sujets et ses commentaires ;
- le temps événementiel, lié au scénario et rythmé par les stimuli et les changements d'affichage de la machine (résultats, warning, etc.) ;
- un temps affectif, rythmé par l'évolution ou les changements d'états du *FoT* du locuteur. Sa particularité tient au fait qu'il n'est pas directement observable, ou alors seulement en partie à travers les signaux physiologiques. Il interroge également sur dans quelle mesure il est lié (ou non) aux deux autres niveaux.

Les deux premiers niveaux forment des axes auxquels il est possible de se référer lors des analyses. C'est ce que nous avons fait indirectement en analysant la répartition des événements par phase du scénario ou la tâche dans laquelle le sujet se trouvait

lors de l'occurrence (relation au temps événementiel), ou la position temporelle de l'événement vocal par rapport aux prises de parole (relation au temps communicatif).

L'objectif global serait de faire émerger des données ces différents niveaux temporels, c'est-à-dire trouver si les différents types d'événements vocaux, ou certains de leurs paramètres intrinsèques sont organisés / produits dans un niveau temporel particulier.

L'hypothèse est que les micro- et macro-organisations temporelles sont fondamentales, soit parce qu'elles sont cohérentes avec l'évolution des états affectifs des sujets, soit parce qu'elles sont révélatrices de ces états (certaines expressions ou micro-expressions ne seront pas directement porteuses d'information, mais l'organisation temporelle de ces icônes le sera) (Carlier et Graaf, 2006).

L'analyse de l'organisation temporelle nous met face à un problème d'alignement : elle nécessite de mettre en relation notre étiquetage audio-visuel avec les différents éléments situationnels. La difficulté rencontrée est que nos seules connaissances *a priori* des éléments susceptibles d'être pertinents proviennent des analyses décrites dans les parties II. et III. de ce chapitre.

Pourtant, l'idéal serait de partir du script du scénario et des événements physiologiques enregistrés, puis de regarder les labels d'auto-annotation, en gardant à l'esprit le profil psychologique du sujet concerné, et ensuite seulement de nous pencher sur notre étiquetage.

## *IV.2. Une réponse possible : l'approche éthologique de la rythmicité*

Dans la langue courante, le terme « rythme » correspond à une « répétition périodique (d'un phénomène de nature physique, auditive ou visuelle) ».[105] Quant aux rythmes ponctuels, ils sont, en biologie du comportement, « des séquences d'événements ponctuels, c'est-à-dire ayant une durée très brève par rapport à l'intervalle qui sépare deux événements. » (Graff, 2008)

La biologie et la psychologie se sont surtout intéressés aux phénomènes temporels pour des événements qui se répètent à intervalles de temps constants ou faiblement variables (les périodes). En effet, la plupart des phénomènes biologiques, y compris comportementaux, possèdent un rythme propre. Celui-ci est souvent périodique en étant souvent calqué sur des variations environnementales comme le rythme circadien (c'est-à-dire l'alternance jour / nuit), ou sur des phénomènes physiologiques tels que la respiration et les battements cardiaques. Toutefois, le rythme d'événements peut être

[105] Définition du TLFI, consulté la dernière fois le 31/01/2011

beaucoup plus irrégulier. Son étude est alors plus complexe, et d'autant plus que de nombreux autres paramètres interviennent souvent (*e.g.* en musique, l'intensité, le timbre, la hauteurs des sons, etc.). Il s'agit donc dans ce cas de réduire l'étude à la seule dimension temporelle. Les événements sont alors considérés comme identiques les uns aux autres (tout-ou-rien), et l'intérêt porte seulement sur la durée des intervalles entre les événements (IPI, pour *Inter-Pulse-Intervals* en anglais). Dans le cas d'événements de durée non négligeable, la référence de l'événement sera sa date de début, et les analyses porteront sur les intervalles entre le début des événements (IOI, pour *Inter-Onset-Intervals* (*ibid*)).

Analyser ces IPIs ou IOIs permet d'étudier à la fois la fréquence (ou plutôt la cadence lorsque les intervalles sont variables), et la régularité des événements considérés. C'est ce couple fréquence / régularité que nous appelons rythmicité.

Pour étudier cette rythmicité, la représentation graphique des IOIs (dans notre cas) peut se faire sur des chronogrammes (durée des IOIs en fonction du temps). La Figure 47 est un exemple de chronogramme issu de nos données, représentant graphiquement la rythmicité des interjections du sujet F_T, sur l'ensemble du scénario.

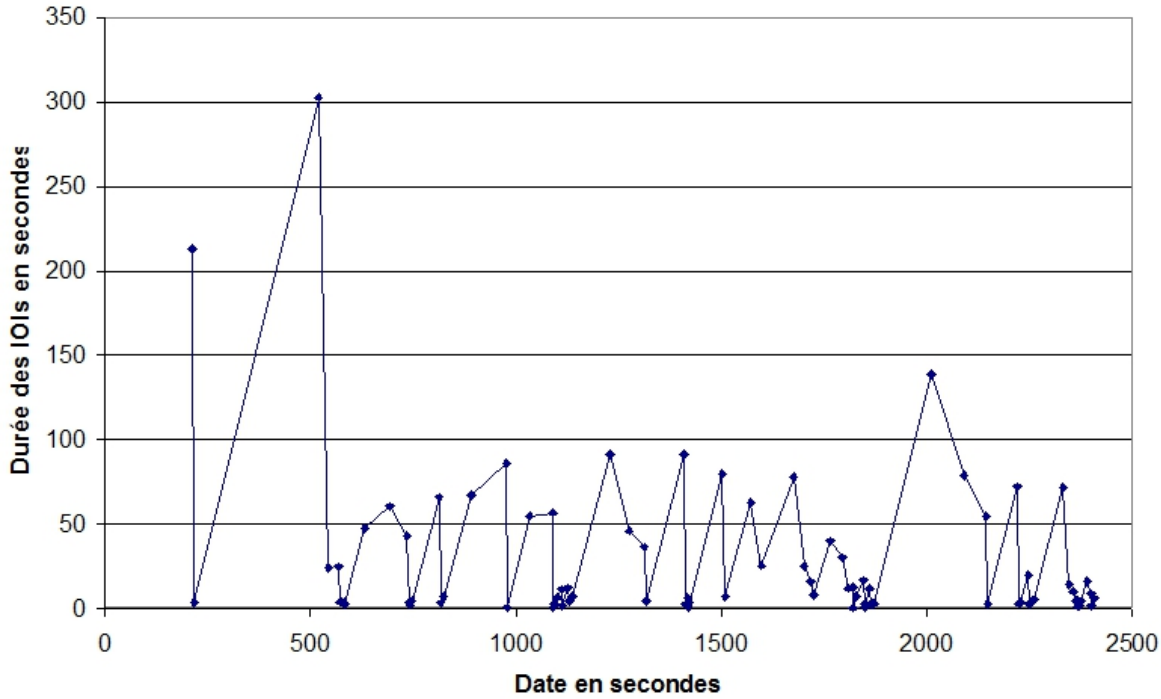



Figure 47: Chronogramme des interjections de F_T

Sur les chronogrammes, il est également souvent utile de représenter la durée des IOIs sur une échelle logarithmique (Figure 48), afin de faire ressortir les proportions, et en même temps de minimiser les valeurs extrêmes des IOIs (et donc ne pas focaliser sur elles -*e.g.* les trois premières valeurs d'IOIs dans notre figure précédente-).

Ces deux chronogrammes (Figures 47 et 48) nous montrent par exemple que les interjections de ce sujet sont le plus souvent groupées temporellement à certains moments du scénario. Il s'agirait ensuite, pour interpréter ces groupements, d'aligner ce chronogramme avec des informations issus des différents niveaux temporels (événementiel, communicatif et émotionnel), afin d'établir des corrélations.

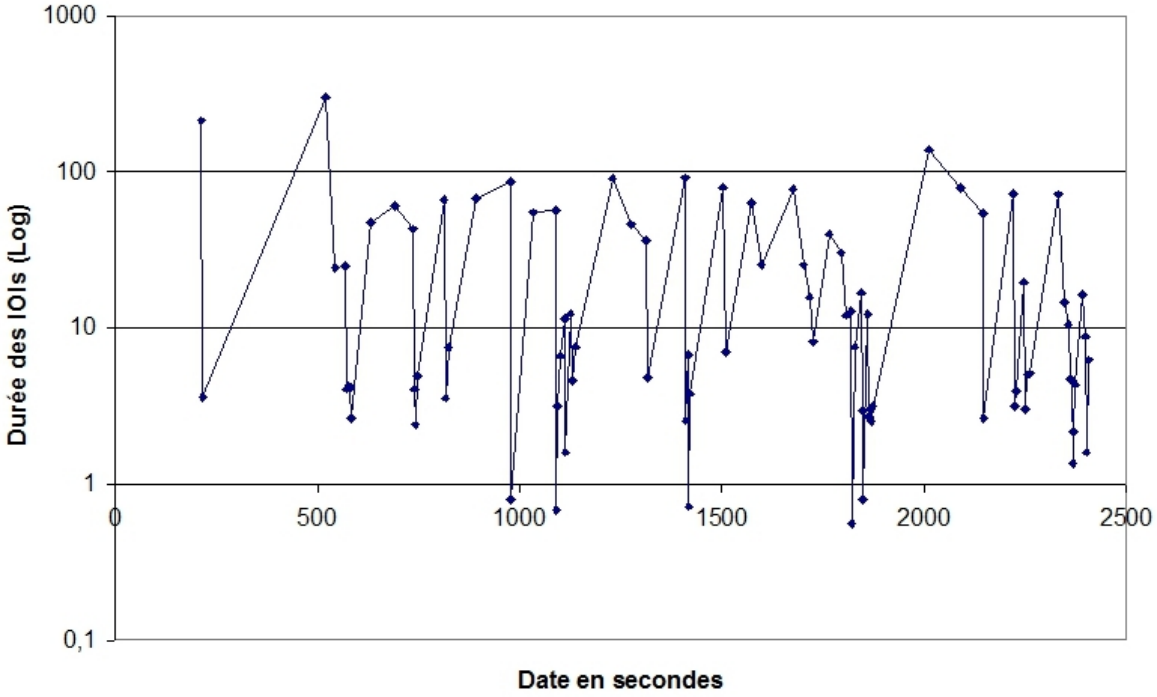



Figure 48: Chronogramme des interjections de F_T avec une échelle logarithmique pour l'axe des ordonnées

Bien entendu, les exemples de chronogrammes proposés ici portent sur les interjections seules, et il serait nécessaire de représenter et comparer des chronogrammes portant sur différents ensembles d'expressions ou types d'expressions (*e.g.* tout événement vocal, uniquement les bruits de bouche ingressifs, uniquement les bruits de bouche « bloqués », uniquement les interjections « complexe », etc.). Procéder ainsi pourrait montrer l'appartenance d'événements vocaux, ou de certains de leurs paramètres, à un niveau temporel particulier.

Par ailleurs, la régularité des comportements est calculable par le biais de l'écart-type des IOIs relevés et par leur coefficient de variation. La médiane permet quant à elle de déterminer la cadence, rapide ou lente, de production du comportement. Représenter ces dernières informations graphiquement (*e.g.* par la répartition des IOIs par durée -Figure 49-) pour différents comportements permet par la suite de comparer leur rythmicité et de mettre en évidence similitudes et différences.

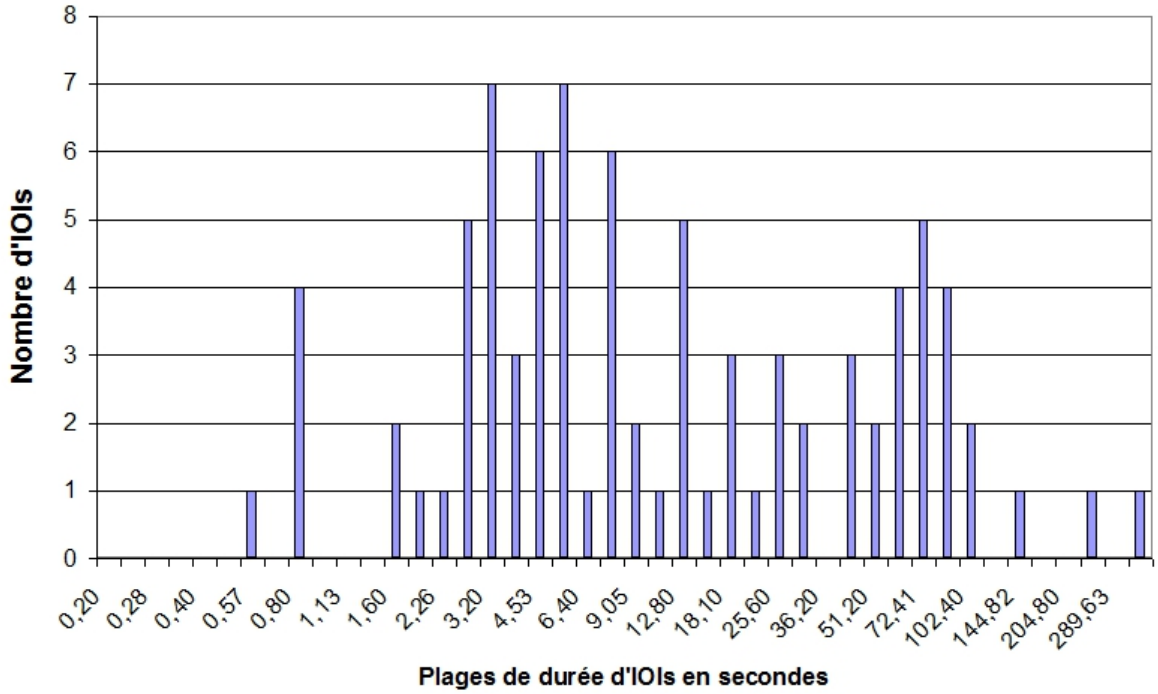



Figure 49: Répartition des IOIs des interjections de F_T par durée

Outre le fait de pouvoir faire émerger des données les différents niveaux temporels éventuels, l'analyse de la rythmicité permettrait de tester l'hypothèse que la rythmicité d'un comportement particulier, ou d'un ensemble de comportements, est porteuse d'information sur les états de *FoT* des sujets. Il reste néanmoins à savoir comment grouper éventuellement ces comportements pour faire émerger des patrons temporels. De plus, il serait intéressant de vérifier l'hypothèse du LaBiCo[106], selon laquelle un comportement régulier est plus saillant qu'un comportement irrégulier (Graff, 2009). Dans ce cas, il s'agirait d'un paramètre important à prendre en compte pour la synthèse expressive.

> « Dans la littérature des sciences comportementales, les analyses du temps et du domaine de fréquence ont souligné les enjeux de la quasi-périodicité dans l'organisation temporelle du comportement expressif [...]. La modélisation du "retard séquentiel" et les HMM qui en sont liées ont été informatifs par rapport à la dynamique des actions discrètes et des états individuels et dyadiques. […] Boker a identifié des "ruptures de symétrie" dans lesquelles le patron d'anticipation entre les partenaires change de manière abrupte. Les erreurs dans la modélisation de ces ruptures pourraient sérieusement compromettre les estimations de l'influence mutuelle. »[107] (Cohn, 2007, p.12)

[106] Laboratoire de Biologie Comportementale de Grenoble

[107] « In behavioral science literature, time- and frequency domain analyses have emphasized issues of quasi-periodicity in the timing of expressive behavior […]. Lag-sequential and related hidden Markov modeling have been informative with respect to the dynamics of discrete actions and individual and dyadic states. […] Boker identified “symmetry breaks”, in which the pattern of lead-lag relationships between partners abruptly shifts. Failure to model these breaks may seriously compromise estimates of mutual influence. »

Comme Cohn l'a suggéré ici, l'analyse des différentes organisations temporelles et de la rythmicité des comportements est une perspective de recherche importante. Les outils proposés par les éthologues apparaissent adaptés pour nous permettre de l'approfondir.

# V. Résumé

Nous avons présenté dans ce chapitre des analyses de l'organisation temporelle des différents événements vocaux étiquetés, en fonction de paramètres situationnels. Il en ressort qu'un grand nombre d'événements vocaux, de nature variée, sont produits dans nos données. Chaque type de bruits de bouche et d'interjections est produit avec une fréquence plus ou moins élevée. Cette fréquence peut être influencée par des paramètres situationnels tels que la phase globale du scénario ou la tâche dans laquelle le sujet se trouve. D'autre part, leur production semble être en relation avec les productions de parole du sujet, mais aussi les productions des autres événements vocaux.

En parallèle, nous avons mis en évidence des variations inter-individuelles dans le comportement vocal de nos sujet. Ainsi, des patrons de comportements particuliers pourraient jouer un rôle dans la caractérisation des personnes. Cela ouvre ainsi des perspectives technologiques en reconnaissance, comme en synthèse audio-visuelle. En somme, les différents événements vocaux sont produits de manière irrégulière, soumis à un certain nombre de contraintes, liées à l'individu ou à la situation, dont dépendent également leur nature et leurs paramètres intrinsèques.

Outre le fait de valider la pertinence de notre étiquetage (notamment de la distinction entre bruits de bouche et interjections), ces observations nous amènent à approfondir le rôle de la rythmicité dans le comportement vocal, ainsi que la notion de niveaux temporels.

Nous nous sommes intéressés ici aux événements vocaux au sein d'une communication temporellement située. Dans le prochain chapitre, nous allons considérer les paramètres intrinsèques, et en particulier prosodiques, de ces mêmes événements, et approfondir leur dimension linguistique. Notons que la composante temporelle reste présente dans nos analyses, puisque la durée d'un événement apparaît comme un paramètre prosodique important.

Le contrôle involontaire du flux respiratoire « bruyant » et de formes sonores est lié à des états physiologiques ou des caractéristiques indiosyncrasiques. Nous faisons l'hypothèse que le contrôle volontaire de la régulation du flux d'air est l'amorce de l'installation du code. À ce contrôle de la durée peut s'ajouter les contrôles d'intensité, de F0 et de qualité de voix, classiquement décrits dans la prosodie. Nous proposons que leur prise en charge volontaire, destinés à produire des formes sonores potentiellement porteuses d'informations communicatives, est le premier pas vers la lexicalisation.

# CHAPITRE 7 : VERS UNE PROSODIE AUDIO-VISUELLE DES MICRO-ÉVÉNEMENTS

Dans le chapitre précédent, nous nous sommes intéressés aux différentes organisations temporelles des événements vocaux, en les considérant comme événements d'une communication située. Nous allons dans ce court chapitre poser les bases de ce qui pourra être ultérieurement une analyse acoustique descriptive et « explicative » de ces signaux. Nous avons dû, étant donnée la lourdeur du travail de fouille, d'étiquetage et d'annotation articulatoire impressionniste de ces micro-gestes vocaux du *FoT*, nous restreindre à un simple inventaire. Ainsi, nous n'avons ici ni montré la pertinence communicative (puisque nous n'avons pas encore pu mener de test perceptif analogue à ceux menés pour les micro-gestes faciaux), ni établi de description acoustique.

Un classement des différents micro-gestes vocaux a été établi sur un continuum correspondant à une complexification du contrôle prosodique (partie II.1.2.). Ce continuum s'étend des événements vocaux qui seraient une conséquence physiologique automatique, ou une expression involontaire, jusqu'aux interjections, dernière étape avant la double-articulation, qui permet l'accès au lexique morpho-syntaxiquement « utilisable ».

Nous n'avons pas encore pu mener de mesures perceptives, ni d'analyses prosodiques (acoustiques et physiologiques[108]) des événements vocaux. Cependant, le travail de Signorello et al. (2010), auquel nous avons participé, a montré une qualité « langagière » des événements que nous supposons volontairement contrôlés dans leur durée (partie II.2., et Signorello, Aubergé, Vanpé, Granjon, & Audibert, 2010). Notons que cette supposition est établie sans mesure objective, uniquement à partir de notre étiquetage impressionniste et surtout à partir de l'écoute.

Cette étude a consisté à tester perceptivement, en audio seul, visuel seul et audio-visuel, des événements vocaux de nos données, présentés aux juges dans un ordre en relation étroite avec le continuum. Le but était, en autres, de trouver à partir de quand, dans notre continuum, les micro-événements audio-visuels apportent des indices concernant la culture et / ou la langue.

Cette expérimentation nous a permis de mieux appréhender la distinction entre indices et signaux, et de concevoir un réajustement de la notion de double-articulation (partie III.).

[108] Étant donnée l'absence de mesures articulatoires autres que la vidéo des visages.

# I. Variabilité intrinsèque des événements vocaux

Nous avons vu Chapitre 6 I.2. qu'il existait des liens résistants entre les types d'événements vocaux et certains des paramètres « intrinsèques » que nous avons notés. Ici, nous allons nous attacher à noter le nombre d'items relevés pour chacun des traits étiquetés. Cela nous permettra ainsi de préciser la variabilité articulatoire et acoustique des événements rencontrés dans notre corpus.

Précisons que nous n'analyserons pas ici la nature des « rires » (étiquetés « expirations brutales » dans nos données), même si nous avons déjà cumulé quelques mesures après Loyau (2007). Il s'agit en effet d'un objet de recherche complexe en soi, qui joue clairement des rôles multiples et fondamentaux du FoT en regard événements vocaux. Leur première particularité est qu'il existe non pas « un rire », mais « des rires », tous différents dans leur nature intrinsèque, leur nature informationnelle, leur nature rythmique, leur fonction ou encore leur lien au contexte (Loyau, 2007 ; Petridis, 2008 ; Urbain et al., 2010). Il serait ainsi impossible d'en traiter tous les aspects ici.

## *I.1. Bruits de bouche*

122 bruits de bouche sur les 2358 (soit 5%) sont voisés. Parmi eux, la majorité (107, soit 88%) ne porte aucune qualité de voix/son particulière (*i.e.* autre que modale). Parmi les 15 occurrences qui en portent une, 6 (soit 40%) sont chuchotées, et 5 (un tiers) sont soupirées. Notons que 73% des occurrences de bruits de bouche voisés avec qualité de voix sont produits avec un flux d'air continu (Tableau 24).

Par ailleurs, il est possible de relever de la « qualité de son » sans pour autant que les bruits de bouche soient voisés[109], pour 35 occurrences. 71% d'entre elles sont soupirées, et les autres se répartissent principalement entre « *creaky* » et « tremblante » (4 occurrences chaque). Comme pour les bruits de bouche voisés, ces occurrences porteuses de qualité de voix/son sont majoritairement produites dans un flux d'air continu (30/35, soit 86% d'entre elles).

Toutefois, la proportion des cas de bruits de bouche non voisés porteurs de qualité de voix/son, par rapport à l'ensemble des événements vocaux relevés, est relativement faible : ces cas concernent seulement 2% (35/2236) des bruits de bouche non voisés (Tableau 24).

[109] Cela illustre et justifie ainsi l'utilisation de l'expression « qualité de son » au lieu de « qualité de voix ».

Tableau 24 : Nombre d'occurrences de bruits de bouche selon leur paramètre de voisement et de qualité de voix / de son, en fonction de leur type de flux d'air

| Type de flux d'air du bruit de bouche --> | Qualité de voix / son | bloqué | gêné | continu | Total |
|---|---|---|---|---|---|
| Non voisés | modale | 1300 | 84 | 817 | **2200** |
| | chuchotée | 1 | | 1 | **2** |
| | creaky | | | 4 | **4** |
| | soupirée | 3 | 1 | 21 | **25** |
| | tremblante | | | 4 | **4** |
| **Total des non voisés** | | 1304 | 85 | 847 | **2236** |
| Voisés | modale | 26 | 32 | 49 | **107** |
| | chuchotée | 2 | 1 | 3 | **6** |
| | creaky | | 1 | 2 | **3** |
| | murmurée | | | 1 | **1** |
| | soupirée | | | 5 | **5** |
| **Total des voisés** | | 28 | 34 | 60 | **122** |
| Total | | **1332** | **119** | **907** | **2358** |

Plus précisément, si nous analysons ces mêmes paramètres en fonction du type articulatoire des bruits de bouche (*cf.* Annexe 11), nous remarquons que les cas de bruits de bouche non voisés possédant une « qualité de son » particulière sont principalement des « expirations » (77% des bruits de bouche non voisés). En effet, ce dernier type de bruit de bouche peut être un support pour toute les qualités de voix/son relevées dans cette étude (sauf « tremblante »), que l'expiration soit voisée ou non. 8% (27/353) des expirations non voisées sont porteuses de qualité de voix/son, ce qui est aussi le cas de 9 des 33 expirations voisées (soit 27%).

D'autre part, il est à noter que seuls certains types articulatoires de bruits de bouche ont été porteurs, au moins une fois dans nos donnés, d'une qualité de voix/son particulière, qu'ils soient ou non voisés : les expirations (36 occurrences sur 386), les gémissements (1/8), les inspirations (6/542), « relâche sa respiration » (4/110), « langue-lèvres mouillées » (1/86), et enfin les occlusions (1/134).

## *I.2. Interjections*

Concernant les interjections, si nous reprenons le Tableau 6 (p.216) en précisant les qualités de voix (Tableau 25), nous pouvons alors dégager des tendances quant à leur utilisation en fonction des types phonétiques d'interjections.

Tableau 25 : Inventaire des interjections relevées et de leur qualité de voix, par ordre de fréquence en fonction du type phonétique.

| Type | Qualité de voix | Nombre d'occurrences | Pourcentage du total |
|---|---|---|---|
| V | modal | 226 | 46,1% |
| V-variantes | chuchoté | 19 | 3,9% |
| | creaky | 40 | 8,2% |
| | murmuré | 5 | 1,0% |
| | nasalisé | 9 | 1,8% |
| | soupiré | 17 | 3,5% |
| Total V-variante | | 90 | 18,4% |
| **Total V** | | **316** | **64,5%** |
| VC | modal | 46 | 9,4% |
| VC-variantes | creaky | 4 | 0,8% |
| | murmuré | 1 | 0,2% |
| | nasalisé | 1 | 0,2% |
| | soupiré | 6 | 1,2% |
| Total VC-variante | | 12 | 2,4% |
| **Total VC** | | **58** | **11,8%** |
| CV | modal | 45 | 9,2% |
| CV-variantes | chuchoté | 1 | 0,2% |
| | creaky | 2 | 0,4% |
| | nasalisé | 3 | 0,6% |
| | soupiré | 1 | 0,2% |
| Total CV-variante | | 7 | 1,4% |
| **Total CV** | | **52** | **10,6%** |
| C | modal | 23 | 4,7% |
| C-variantes | chuchoté | 4 | 0,8% |
| | creaky | 4 | 0,8% |
| | soupiré | 3 | 0,6% |
| Total C-variante | | 11 | 2,2% |
| **Total C** | | **34** | **6,9%** |
| Comb | modal | 21 | 4,3% |
| Comb-variantes | chuchoté | 1 | 0,2% |
| | creaky | 1 | 0,2% |
| | murmuré | 3 | 0,6% |
| | nasalisé | 2 | 0,4% |
| | soupiré | 2 | 0,4% |
| Total Comb-variante | | 9 | 1,8% |
| **Total Comb** | | **30** | **6,1%** |
| Total | | 490 | 100,0% |

Ainsi, seules les qualités de voix *creaky* et soupiré ont été rencontrées comme attribut de chacun des types phonétiques d'interjection.

Plus précisément, notons que pour les variantes d'interjections, le type phonétique semble influencer la qualité de voix utilisée. En effet, bien que cela reste à montrer statistiquement avec des données moins restreintes, il semblerait que les qualités de voix les plus récurrentes soient :

- creaky pour V : 44% (40/90) des V-variantes sont « creaky » ;
- soupiré pour VC : 50% (6/12) des VC-variantes sont soupirées ;
- nasalisé pour CV : 43% (3/7) des CV-variantes sont nasalisées ;
- chuchoté et creaky pour C : 36% (4/11) des C-variantes sont chuchotées, et la même proportion est « creaky » ;
- murmuré pour Comb : un tiers (3/9) des Comb-variantes sont murmurées.

De plus, les différentes qualités de voix sont plus ou moins fréquentes (voir Tableau 25) : alors que la voix « creaky » concerne 40% de toutes les variantes d'interjections, la voix murmurée n'en concerne que 7%.

Tableau 26: Nombre d'occurrences et pourcentage du total des variantes, en fonction de la qualité de voix.

| Qualité de voix | Nombre d'occurrences | Pourcentage sur le nombre total de variantes |
|---|---|---|
| creaky | 51 | 40% |
| soupiré | 29 | 22% |
| chuchoté | 25 | 19% |
| nasalisé | 15 | 12% |
| murmuré | 9 | 7% |

Plus précisément, les différentes qualités de voix ne concernent pas toutes les mêmes types phonétiques d'interjections :

- 76% (19/25) des interjections chuchotées sont purement vocaliques (V), et 16% (4/25) sont purement consonantiques (C) ;
- 78% (40/51) des interjections « creaky » sont vocaliques (V). Elles représentent 13% des V total, et 44% des V-variantes ;
- un tiers (3/9) des interjections murmurées sont des combinaisons (Comb), ce qui représente également 1/3 des Comb-variantes ;
- 20% (3/15) des interjections nasalisées sont des VC, ce qui représente 43% (3/7) des VC-variantes ;
- 59% (17/29) des interjections soupirées sont purement vocaliques (ce qui représente 19% (17/90) des V-variantes), puis 21% (6/29) sont des VC.

Les différentes qualités de voix semblent donc inter-dépendantes des types phonétiques d'interjections produits.

Ces analyses des interjections interrogent sur les qualités vocaliques utilisées et sur leur statut phonologique : les types de voyelles produites sont-elles liées à un effort vocalique minimal ? Sont-elles éventuellement explicables physiologiquement ? Le [Ø :] est-il un contrôle spécifique du français ? Et d'un point de vue phonologique, « mmh » est-il une variante phonologique de « euh », de même qu'un *trill* bilabial vis-à-vis de « pff », ou encore « chchch » comme variante de « chut » ?

# II. Du contrôle prosodique au langage

Il n'est question ici ni des différentes définitions, descriptions phonétiques, ou modélisations phonologiques, ni des faiblesses des mesures acoustiques ou de la méconnaissance articulatoire de la prosodie (Martin, 2006 ; Rossi, 1999 ; Lacheret-Dujour & Beaugendre, 1999).

Il est trivial de rappeler que la prosodie est un vecteur privilégié pour l'expression par la parole des états affectifs (*e.g.* Scherer, Ladd, & Silverman, 1984, ou Aubergé 2002a). Un consensus est établi sur trois paramètres acoustiques, sur lesquels se construirait la prosodie : la fréquence fondamentale (F0), l'intensité et la durée. Il est à noter que la durée n'est pas un paramètre du signal, mais une mesure à établir sur l'axe temporel, si tant est – et c'est un point central non résolu – que soient définis un ou des « segments » à mesurer. De plus, les paramètres spectraux du timbre de la voix, et plus généralement de la qualité de voix, sont ajoutés aux trois paramètres, en particulier par les recherches qui accordent à la prosodie la fonction idiolectale ou surtout la fonction émotionnelle (Aubergé, 2002a).

Rappelons simplement que caractériser la prosodie par ces paramètres ramène à une ambiguïté fonctionnelle : la phonologie des segments, par la paire minimale, montre sa pertinence pour l'accès lexical ; F0, durée et intensité (et plus ou moins clairement la qualité de voix pour des langues comme le vietnamien) actionnent également la fonction d'accès lexical avec les tons ou les accents lexicaux. Les autres fonctions supra-lexicales, linguistiques et pragmatiques jouées par la prosodie sur ces mêmes paramètres sont multiples (Aubergé, 2002b ; Martin, 2011). Sans aucunement entrer dans un débat, ce que nous appelons prosodie pour les bruits de bouche, donc sans segment phonologique, est le matériel acoustique même du son. La durée en est le paramètre clé, et la qualité de son est potentiellement une « genèse » de la qualité de voyelle. Notons que dans nos éléments « primitifs », la durée pose peu le problème de la détermination du segment à mesurer : de nombreux bruits de bouche sont ancrés sur une phase respiratoire, quant aux plus complexes, ils ne dépassent guère ce qui serait appelé syllabe pour une entité lexicalisée. Ainsi, au-delà de la durée, le problème du rythme de ces énoncés (complets puisque isolés) se présente également sous un matériel très simple.

## *II.1.Tentative impressionniste de qualification du contrôle*

### II.1.1. Traces perceptibles du contrôle prosodique

Certains événements vocaux sont la conséquence d'une fonction végétative (telle la déglutition), ou une conséquence physiologique, par exemple en lien avec l'état émotionnel du sujet (*cf.* l'effet *push* de Scherer -Chapitre 1 II.3.-).

Le contrôle[110] de la respiration, à travers celui de l'inspiration ou de l'expiration, permet de contrôler la durée de la phase respiratoire. Rappelons que, selon notre point de vue, c'est la prise de contrôle volontaire *vs.* geste involontaire qui nous intéresse en tant que facteur de discrimination d'un statut non langagier *vs.* langagier, cela quelque soit la pertinence communicative des événements. Il pourrait s'agir du premier pas vers une sorte de « prosodie », destinée à être entendue ou non. Notons que ce contrôle de la respiration peut avoir pour fonction de modifier l'état somatique / physiologique.

La respiration est un rythme biologique fondamental pour l'humain : il s'agit d'une contrainte dans les productions des micro-événements audibles. Le simple contrôle de la respiration a une influence sur la durée des événements. Ainsi, au niveau perceptif, la durée est certainement un des premiers paramètres qui manifeste un contrôle volontaire.

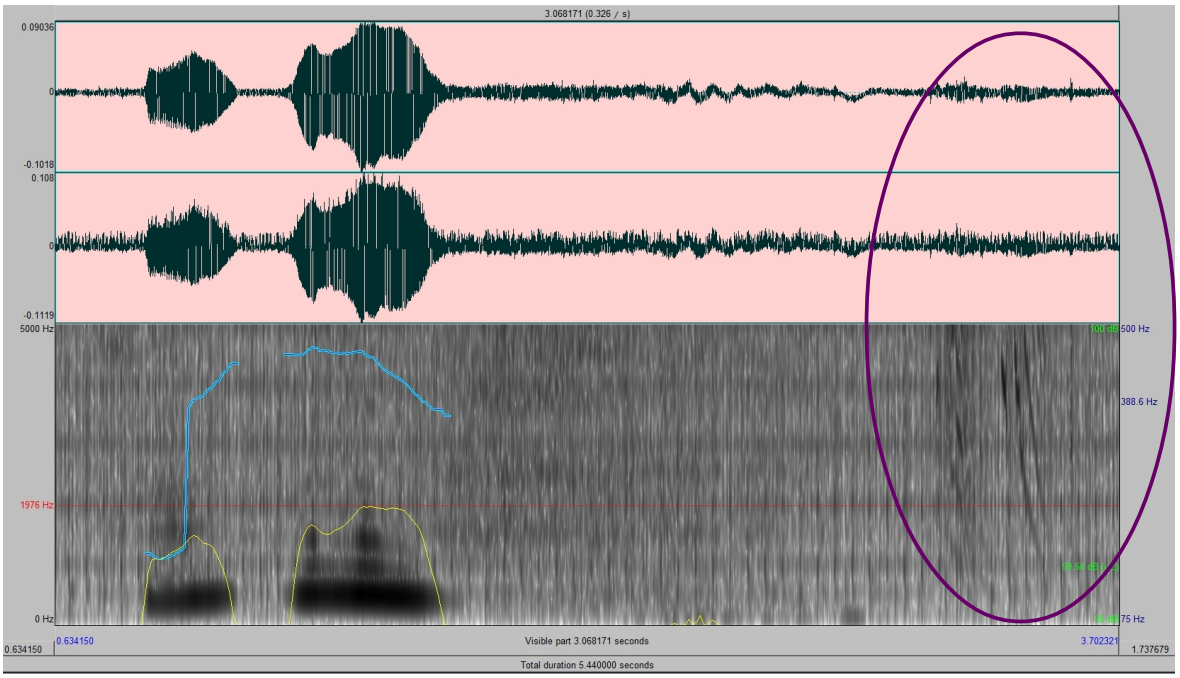



Figure 50: Spectrogramme d'un gémissement (avec variation de la F0, en bleu) suivi d'une friction bilabiale ingressive (indiquée par l'ellipse violette) - sujet F_T -

[110] Le contrôle dont nous parlons peut-être un contrôle volontaire ou involontaire, au sens d'Aubergé (2002).

La présence de voisement, avec la variation éventuelle de la F0 (Figure 50), et la présence d'une qualité de voix particulière, sont des paramètres rencontrées dans nos données. Notons que plusieurs études ont montré l'importance de la qualité de voix pour l'expression émotionnelle du locuteur dans la parole (entre autre Scherer et collègues).

Nous cherchons à distinguer, parmi les événements vocaux relevés, ceux dont la durée peut être contrôlée, et ceux dont la durée est la conséquence des gestes automatiquement déclenchés. Les événements vocaux sont ainsi vus comme un contrôle articulatoire du conduit vocal dans le temps, avec un contrôle de la durée, ou une durée en tant que paramètre « par défaut ».

Au niveau des événements vocaux de type interjection, leur nature phonétique, c'est-à-dire le fait qu'ils soient composés de phonème(s), implique un contrôle volontaire dont la prosodie est « classiquement » analysable.

### II.1.2. De la conséquence physiologique au langage par la prise de contrôle : premier classement

Comme nous l'avons évoqué Chapitre 5 I.1.5., un de nos objectifs est d'ordonner tous ces événements vocaux, ces items non-lexicaux (de la « deuxième articulation » selon Martinet (1964)) sur un continuum, un axe orienté, où leur distance à la « première articulation », linguistique, diminue. Notre hypothèse est qu'en parallèle de cette diminution, le contrôle des paramètres prosodiques augmenterait : au début de l'axe les bruits de bouche en tant que conséquence physiologique ou musculaire (*e.g.* les « relâchement d'articulateurs ») ; puis éventuellement les bruits comme « qualité de sons » par eux-mêmes ; ensuite les items dont la durée est contrôlée ; enfin ceux dont le contrôle s'étend à la F0, l'intensité et / ou contrôle la qualité de voix.

Ainsi, nous avons cherché à classer, par expertise, des événements vocaux de chacun des types relevés, en fonction de leur degré de contrôle. Pour cela, nous sommes partis d'une sélection d'événements effectivement rencontrés dans nos données, puisque chaque occurrence est unique, et souvent contrôlée de manière différente. En cohérence avec les réflexions du début de la partie II.1., nos critères ont été en premier lieu, pour les bruits de bouche, le contrôle de la durée (notamment par le contrôle du flux d'air et de son intensité), puis ceux de la « qualité de son » et du voisement (ainsi que sa variation). Pour les interjections, nous avons considéré la complexité de leur type phonétique ainsi que le contrôle de leur durée et la présence de qualité de voix.

À partir de nos différents types articulatoires et phonétiques d'événements vocaux, nous avons établi 15 niveaux de contrôle sur le continuum, qui servent de base à la classification d'occurrences.

1) « relâchement d'articulateur(s) » ;
2) « déglutit » et « relâchement de la respiration » (notons que l'expiration qui suit l'occlusion glottale peut avoir sa durée contrôlée par le sujet) ;
3) occlusion et « interaction langue-lèvres » ;
4) inspiration et « occlusion ingressive_inspiration » ;
5) click, qui implique un contrôle volontaire de sa production ;
6) expiration sans contrôle prosodique ;
7) raclement de gorge (notons qu'à ce niveau du continuum, il s'agit d'un raclement de gorge volontaire, et non une simple réaction physiologique à une gêne) ;
8) friction ;
9) expiration avec contrôle de la durée, et éventuellement du voisement et / ou de la qualité de voix ;
10) gémissement, qui implique un voisement, et éventuellement une variation de la F0 et / ou de la qualité de voix ;
11) interjection consonantique, « C » ;
12) interjection vocalique, « V » ;
13) interjection de type « VC » ;
14) interjection de type « CV » ;
15) interjection « complexe », de type « Comb ».

D'une manière globale :

- les 6 premiers niveaux relèvent de sons non phonétiques et supposés sans contrôle prosodique volontaire ;
- du 7ème au 10ème niveau se trouvent des bruits de bouche non phonologiques pour le français, mais contrôlés prosodiquement. Ce contrôle peut concerner la durée de l'item (soit par l'allongement de la durée du flux d'air, soit par le réduplication d'un élément), sa « qualité de son » ou encore la variation de sa F0 ;
- du 11ème au 15ème niveau sont classées les interjections, items composés de phonèmes de la langue et pré-lexicaux.

Des niveaux de contrôle intermédiaires sont ensuite ajouté pour tenir compte des spécificités de chaque occurrence d'événement (en particulier de leur niveau de contrôle prosodique dont les paramètres sont décrits ci-dessus.

Il est à noter que cet ordre est indicatif : il n'est pas strict et doit être adapté à chaque ensemble particulier d'occurrences. Il n'est pas lié à un facteur articulatoire, mais à un facteur de contrôle prosodique (une sorte de « manière articulatoire »). C'est pour cette raison que cet ordre peut ne pas être le même pour les différents sujets : le sujet sait la plupart du temps contrôler la durée et la prosodie des différents événements, mais va ou non utiliser ce contrôle sur chacun des types d'occurrences (*cf.* la variation

inter-sujets observée Chapitre 6 III.). Par exemple M_J a une tendance à contrôler particulièrement ses inspirations et expirations, y compris en ajoutant du voisement, voire de la qualité de voix. C'est pourquoi au niveau du contrôle, ses expirations se rapprochent plus des interjections que celles des autres sujets.

## *II.2. La discrimination perceptive audio-visuelle de la langue / culture*

Cette partie résume une expérimentation menée par (Signorello, Véronique Aubergé, Vanpé, Granjon, & Nicolas Audibert, 2010), qui a permis de tester la classification présentée dans la partie précédente, d'un point de vue langagier. Plus précisément, elle a permis de tester l'hypothèse que la nature langagière des événements sonores est construite par le contrôle de la prosodie, et que les premiers faits de langage apparaissent ainsi bien avant que ces sons ne soient doublement articulés.

L'objectif de cette expérimentation a été de vérifier s'il était possible de discriminer deux langues / cultures, mêmes proches, à travers des micro-événements audio-visuels, non langagiers relevés dans nos données (Signorello, Aubergé, Vanpé, Granjon, & Audibert, 2010), c'est-à-dire :

> « […] de mettre en évidence cette possible identité langagière (ou au minima culturelle) de certains des micro-événements, et de surcroît, mesurer perceptivement lesquels commencent à contenir de telles informations vs. ceux qui n'ont pas de statut langagier. » (*ibid*, p.2)

Elle a consisté en une identification perceptive de la langue / culture, et le degré de confiance accordé à ce choix, par des juges français et italiens[111]. Les jugements ont été réalisés sur des micro-événements acoustiques et visuels, non langagiers, de nos six sujets français. Ils ont de plus été réalisés à partir de trois modalités : Audio seul, Visuel seul ou Audio-Visuel.

Ces événements ont été présentés selon l'ordre déterminé par le continuum (cf. partie II.1.2.), c'est-à-dire selon un contrôle croissant de la prosodie, à partir d'un non contrôle. Chaque sujet a fait l'objet d'une session particulière du test, lors de laquelle l'ordre a été respecté. En effet, l'hypothèse sous-jacente est que le paramètre de contrôle prosodique est informatif, et cela de manière croissante sur notre continuum (*e.g.* les événements de niveau 2 seront moins informatifs que les événements de

[111] « Nous avons retenu, comme paire à contraster, le français et l'italien, de typologie linguistique très proche et culturellement très familières. Ces deux langues / cultures font partie de la même aire euro-méditerranéenne. En outre, il existe une opposition actuelle et claire des langues nationales (français vs. Italien) » (Signorello, Aubergé, Vanpé, Granjon, & Audibert, 2010, p.2)

niveau 5). Il est donc supposé un apprentissage au fur et à mesure de la présentation des stimuli[112].

Les résultats montrent globalement une tendance des juges, des deux nationalités, à identifier correctement les sujets (*i.e*. comme français). Plus précisément, la modalité de présentation a une influence sur cette identification :

> « Toutefois, alors qu'en Visuel seul le jugement est surtout fondé sur l'aspect des sujets (choix des juges invariable dès le début de l'expérimentation), en Audio seul et en Audio-Visuel l'apparition du contrôle prosodique correspond à un "point de stabilité" dans les choix langue / culture. » (*ibid*, p.1)

En Audio et Audio-Visuel, ce point de stabilité dans la détermination de la langue/culture du sujet perçu (c'est-à-dire le moment à partir duquel le juge ne modifie plus son choix, lié dans cette modalité à une augmentation importante du degré de confiance), correspond, sur notre continuum, à l'apparition des stimuli donnés comme contrôlés prosodiquement[113]. Afin de préciser avec quel type d'événements apparaît ce point de stabilité, et donc l'information langagière / culturelle, il serait utile de réitérer l'expérimentation, en focalisant le choix des stimuli sur la partie du continuum concernée.

Quoi qu'il en soit, cette expérimentation a donc permis globalement « de considérer comme "valide" l'interrogation sur l'informativité langagière / culturelle de ces événements subtils, *a priori* non linguistiques » (*ibid*, p.7).

[112] C'est aussi pour éviter cet apprentissage entre les différentes modalités et pour un même sujet, que les juges n'ont passé le test que dans une seule modalité.

[113] Pour des résultats plus précis, se reporter à l'article pré-cité (*ibid*).

# III. Discussion

## III.1. Mesurer la pertinence communicative

Concernant la pertinence communicative, la question de fond qui nous intéresse, au-delà de mesurer si ces micro-gestes apporte ou non de l'information, est de mesurer si l'information résulte d'un contrôle involontaire (état physiologique, émotionnel, humeur, idiosyncrasie) ou d'un contrôle volontaire. Cependant, la nature de l'information véhiculée ne renseigne pas forcément sur la nature du contrôle. Pour étudier cela, il sera nécessaire de mesurer perceptivement les événements vocaux dans des tests ultérieurs, en les associant à leur auto-annotation, comme nous l'avons fait pour les événements de la modalité visuelle. Ainsi, si par exemple la pertinence de l'expression correspondant à l'auto-annotation « exaspéré, j'en ai marre » est vérifiée, nous pourrons supposer qu'il s'agit d'un contrôle volontaire intentionnel. Dans d'autres cas, nous pouvons aussi faire l'hypothèse que la tension d'un sujet stressé ou concentré sera suivi par un relâchement (notamment musculaire), contrôlé involontairement, et qui pourrait être perçu à travers un bruit de bouche du type « relâche sa respiration ». Enfin, concernant la perception d'un événement vocal révélant par exemple un état de fatigue, il s'agira de déterminer si ce micro-geste est déclenché automatiquement (soit comme indice interprétable, soit comme signal de l'état physiologique - la fatigue - du sujet), ou s'il est contrôlé volontairement, dans le cas où le sujet souhaite « faire savoir », intentionnellement, qu'il est fatigué.

Nous reprenons ici l'hypothèse forte d'Aubergé sur la nature du contrôle volontaire-involontaire (1998-2007) pour les micro-gestes du *FoT.* Elle illustre cette hypothèse par l'opposition entre émotions et affects sociaux, qui serait une clé de la compétence langagière (*cf.* Chapitre 2 II.3.2.). En effet, selon elle, l'expression d'une émotion, par nature involontaire, serait un acte de communication, alors que l'expression d'une émotion simulée (*cf.* Damasio, 1994, et Chapitre 1 III.3., notamment Figure 3*)*, et par extension l'ensemble des attitudes intentionnelles construites par la langue, seraient des actes de langage. Les formes de ces deux voies seraient identiques pour les valeurs immédiates *vs.* simulées, mais tandis que l'ancrage temporel serait événementiel pour le contrôle involontaire, il serait linguistique pour le contrôle volontaire.

## *III.2. De l'indice au signal*

À terme, notre ambition est de distinguer les indices des signaux, c'est-à-dire de ce qui est pertinent en soi d'un point de vue communicatif.

Jusqu'à présent, seule la perception inter-culturelle testée dans l'expérimentation de Signorello et al. (2010 - *cf.* partie II.2. -) nous permet, indirectement, de savoir que certains des événements vocaux testés sont des signaux. Il s'agit des événements du classement présenté, à partir desquels les juges ont perçu la langue de nos sujets-locuteurs correctement, et en indiquant un degré de confiance élevé quant à cette perception. À partir du moment où un événement vocal apporte une information linguistique (au sens de « faisant partie d'une langue particulière »), nous pouvons considérer qu'il s'agit d'un signal.

À ce stade de nos analyses, nous ne pouvons aller plus loin quant au statut communicatif des autres événements étudiés.

Les analyses de la partie I. de ce chapitre ont montré que des expirations pouvaient porter de la qualité sonore, ou encore qu'un gémissement pouvait avoir une modulation de sa F0, sans qu'il y ait présence de phonème. Ainsi, certains événements vocaux que nous avons relevés pourraient être considérés comme construits de « prosodie pure », c'est-à-dire d'une prosodie qui n'a pas besoin d'un substrat phonologique d'ancrage.

À terme, il s'agira donc de comprendre, pour ces items de type « pre-phone », si une « qualité de sons » pourrait être initialisée primitivement par le contrôle de la qualité de voix, et si elle peut être liée au *sound symbolism* (*cf.* Chapitre 5 I.1.3.) et à la physiologie pour les expressions émotionnelles (Scherer & Zei, 1989 ; Fredrickson & Levenson, 1998).

# CONCLUSION ET PERSPECTIVES :

## UNE MÉTHODOLOGIE ADAPTÉE À L'ÉTUDE ET À LA MODÉLISATION

## DE LA VARIATION DES MICRO-ÉVÉNEMENTS AUDIO-VISUELS DU *FoT*

## I. De l'étude du comportement expressif de l'humain…

En observant notre corpus d'interaction IHM « écologique », nous avons constaté que les expressions produites étaient loin d'être réduites aux seules expressions involontaires et à la seule fonction émotionnelle. Même dans cette situation très restreinte d'un point de vue de l'interaction sociale, nous avons relevé des signaux sociaux émis par les sujets[114]. Ainsi, des états variés et mélangés, tels que des états mentaux, attitudinaux ou encore affectifs, ont été exprimés par nos sujets, y compris en dehors de leur production de parole. Nous utilisons le terme générique *Feeling of Thinking* pour les nommer.

La variété de ces états et la complexité de leur mélange sont remarquables dans les auto-annotations naïves du corpus, effectuées par les sujets eux-mêmes. Quant à la pertinence de leur utilisation elle a été en partie confirmée dans notre travail, par la bonne reconnaissance globale de ses labels au cours des évaluations perceptives de certaines de nos icônes gestuelles.

Nous avons adopté une approche multimodale, puisque nous faisons globalement l'hypothèse que l'étude des événements de la modalité acoustique dans leur contexte (micro-)gestuel, et l'observation des (micro-)expressions faciales / gestuelles dans leur contexte acoustique, donnera une large échelle d'interprétation en relation avec les paramètres du contexte. Nous avons cependant éludé ici le problème de la multimodalité en observant des icônes gestuelles non sonores, et des événements vocaux, dont nous n'avons pas étudié l'apport visuel, si ce n'est pour les qualifier articulatoirement.

En vue de l'étude des phénomènes du *FoT*, nous avons d'abord étiqueté des expressions et micro-expressions visibles et audibles selon une méthodologie inspirée de l'éthologie, c'est-à-dire de manière naïve et sans *a priori.* De cette manière, nous avons pu nous affranchir des modèles pré-existant de gestualité et d'expressions faciales, et du filtre perceptif qu'ils nous imposent. Étant donnés le corpus

[114] pour une vue d'ensemble des études concernant ces signaux, voir Vinciarelli, Pantic, & Bourlard, 2009

d'expressions spontanées E-Wiz SoundTeacher sur lequel nous avons travaillé et la méthodologie utilisée, cet étiquetage forme une base neutre (du point de vue de l'influence des modèles), de comportements spontanée. Il est à noter que cette base, de par les éléments étiquetés et leur précision, a permis de mettre en inter-relation : les événements eux-mêmes et leurs paramètres, cela entre et au sein des différentes modalités ; les événements du temps communicatifs (*i.e.* les productions de parole du sujet) ; les éléments liés à la situation et au temps événementiel (*e.g.* les stimuli envoyés par la machine au sujet).

De là, nous avons exploré des pistes de recherche liées à la problématique des expressions de *FoT*, c'est-à-dire concernant *:* leur nature, leur nature informationnelle, la manière dont elles sont perçues, leur organisation temporelle, ou encore leur organisation au sein et entre les différentes modalités.

Nous avons d'abord montré l'importance du détail, et ainsi de l'étude des micro-expressions. Nous avons ensuite cherché à poser les prémisses de leur compétence communicationnelle.

Ainsi, notre travail a ensuite consisté à approfondir quels sont les traits morphologiques pertinents des manifestations acoustiques et visuelles, et quels indices communicatifs sont véhiculés par ces morphologies Nous nous sommes également intéressés en parallèle aux paramètres temporels de ces manifestations (*i.e.* à leur dynamique et leur durée, ainsi qu'à leur organisation temporelle), et à l'influence des paramètres situationnels.

D'autre part, un des phénomènes observés parmi les plus marquants est la variabilité inter-sujets dans les occurrences d'événements vocaux, à la fois dans la manière de les utiliser et dans le choix de leur nature. Elle démontre à la fois l'importance du sujet et la complexité à situer le sens de ces événements en dehors du contexte. Les analyses montrent ainsi que ces événements apportent de l'information pertinente concernant le comportement du sujet, en fonction des différentes organisations temporelles de l'interaction (*e.g.* les phases du scénario ou l'organisation de l'interaction en tours de parole). Rappelons que des travaux précédents (Loyau, 2007) et actuels (Petridis, 2008) concernant les rires, montrent que ces derniers sont variables acoustiquement et peuvent être reliés perceptivement à leur paramètres contextuels variés (comportement du locuteur, indices pragmatiques de l'organisation temporelle de l'interaction, états mentaux et affectifs).

Concernant la variation inter-sujets observée, la relation entre les expressions du *FoT* et la personnalité est entrepris à travers l'étude de la stratégie communicative adoptée

par les sujets. Toutefois, la nature des items produits, en fonction du *FoT* et de la personnalité, restent à étudier finement.

Enfin, une des questions posées par cette thèse concerne le statut communicatif des différents phénomènes expressifs multimodaux, c'est-à-dire s'il s'agit d'indices ou de signaux. Nous avons essentiellement pu montrer que les événements vocaux sont susceptibles de porter de l'information culturelle/langagière (et par conséquent d'être des signaux), à partir du moment où ils sont contrôlés prosodiquement, et sans faire partie nécessairement des interjections prélexicales. En parallèle, nous avions fait l'hypothèse de l'existence d'un continuum sur lequel il serait possible de classer les événements vocaux selon l'augmentation du contrôle (en particulier prosodique), dont ils font l'objet, et ainsi la diminution de leur distance aux éléments lexicaux. Nos analyses, ainsi que les résultats de l'expérimentation présentée Chapitre 7 II.2., apparaissent être en cohérence avec cette hypothèse, et ont tendance à valider le continuum proposé. Afin d'approfondir cette dernière hypothèse, il s'agit désormais de préciser la notion de contrôle prosodique, par le biais d'analyses acoustiques et d'expérimentation, et de considérer la nature audio-visuelle des micro-événements.

Étant donnés les résultats préliminaires obtenus, qui incitent à poursuivre dans l'étude de ces phénomènes et paramètres, nous pouvons donc considérer la proposition méthodologique de cette thèse, ainsi que les différentes pistes de recherche abordées, comme pertinente pour l'étude des phénomènes du comportement expressif.

Toutefois, pour poursuivre l'étude des micro-événements vocaux, et en particulier de leur contrôle prosodique, il nous faudra trouver des outils beaucoup plus « subtils » dans l'analyse du signal, et beaucoup plus paramétrables que ceux jusqu'alors utilisés dans cette étude préliminaire (*PRAAT* pour l'acoustique)[115]. En effet, les événements vocaux sont de natures peu typiques pour la phonétique (*cf.* un exemple de spectrogramme Figure 50, p.252). Cela rejoint certainement les grands problèmes posés par les analyses du signal de parole, et leurs pièges récurrents (Martin, 2008).

[115] L'utilisation du logiciel Winpitch (Martin, 2004), un équivalent du logiciel PRAAT, pourrait être envisagée. Winpitch possède l'avantage d'intégrer une fonction de resynthèse de qualité, qui nous permettrait de vérifier la pertinence des paramètres acoustiques que nous extrairons de nos signaux, et qui seront supposés caractéristiques d'un type particulier d'événements. De plus, la version LTL de Winpitch, destiné à l'enseignement de la prosodie, permet la « visualisation des facteurs prosodiques en temps réel » (*ibid*, p.71), en ayant avec le flux vidéo synchronisé et visible en parallèle (ce qui est impossible avec le logiciel PRAAT). Puisque notre approche se veut multimodale, et que nos dernières analyses s'attachent à étudier la prosodie audio-visuelle des micro-événements, ce logiciel apparaît donc particulièrement adapté pour notre étude.

# II. … à sa modélisation et ses perspectives applicatives

Nous avons vu Chapitre 3 III.3., qu'il est à terme indispensable de tester les icônes relatives à l'expression du *FoT* dans des agents virtuels aptes à synthétiser des paramètres comme la dynamique des mouvements. L'un des plus aboutis globalement à ce jour pour des gestes comme ceux que nous avons identifiés étant l'agent virtuel expressif paramétrable GRETA, développé par Pelachaud et collègues. Un autre avantage de cet agent est d'être adaptatif.

Ces synthèses de nos icônes pourraient permettre une évaluation directe du comportement expressif modélisé, en relation aux performances du naturel, en même temps qu'une mesure d'une éventuelle valeur ajoutée aux modèles déjà bien étayés.

Les éléments et paramètres à prendre en compte lors de la modélisation d'un agent communicant expressif sont nombreux et variés, comme nous l'avons montré à la suite de travaux tels ceux du réseau Humaine ou de Martin, Niewiadomski, Devillers, Buisine, & Pelachaud (2006). Ainsi, les recherches en cours tentent constamment d'améliorer la crédibilité et le réalisme de ces agents, en les dotant de capacités de plus en plus nombreuses (*e.g.* la production de *backchannel* multimodal en dehors de ses tours de parole (Heylen, 2007) ou encore d'une personnalité (*e.g.* Egges, Kshirsagar, & Magnenat-Thalmann, 2004 ; Mancini, 2008).

Dans ce but, la variabilité inter-sujets dans leur comportement doit être prise en considération dans les modélisations. Au préalable, il est nécessaire de montrer que cette variabilité rencontrée dans les comportements des sujets (testés et validés en reconnaissance sur des juges naïfs pour les modalités visuelles) nous permettent d'établir des groupes de sujets. Il s'agit par la suite de modéliser pour chacun de ces groupes de sujets, ayant une cohérence dans leur comportement, la manière de produire ce comportement, en fonction notamment des éléments objectifs de la situation.

Par exemple si nous générons par la suite un agent avec un type de comportement gestuel *(e.g.* de nombreux gestes à faible excursion), il faudra également que ses autres caractéristiques comportementales s'inscrivent dans la même cohérence (*e.g.* qu'il produise peu d'interjections mais très perceptibles et contrôlés, et surtout pendant les phases d'induction). En effet, lorsque le comportement d'un agent est modélisé, l'objectif n'est pas qu'il ait un comportement, une personnalité correspondant à une moyenne, un standard (universel ou non). Ce qui est essentiel, c'est que ce comportement, et surtout la manière dont il est perçue par autrui (ce qui est

généralement appelé « personnalité » lorsqu'il est considéré dans son ensemble), soit cohérent dans la nature même des éléments qui le composent, et surtout dans les différentes organisations de ces éléments. À plus long terme il s'agira aussi de déterminer les groupes de sujets les mieux perçus (c'est-à-dire les états exprimés les mieux reconnus et les mieux acceptés) et dont il est préférable de modéliser le comportement sur des agents.

En outre, ce comportement modélisé doit être adapté au contexte dans toute sa complexité (au type d'interaction dans le cas des agents communicant), afin d'être accepté par le plus grand nombre d'utilisateurs. Dans notre cas, étant donnés nos sujets, le comportement qui pourra inspirer une modélisation au niveau gestuel / facial sera ainsi un comportement féminin et lié à la culture française, bien que certains phénomènes que nous avons relevés seront plus génériques.

# III. Conclusion générale

Nous nous attendions à ce que les sujets aient des comportements différents, et nous gardions à l'esprit qu'indices idiosyncrasiques et indices objectifs (voire signaux), allaient se mélanger dans leurs expressions. Cela s'est effectivement vérifié.

Au sein d'une approche multimodale (face, gestes, événements vocaux), nous avons cherché à travers quelle(s) modalité(s), et dans quelle(s) temporalité(s), sont véhiculées les différentes informations concernant les états de *FoT* de nos sujets. Globalement, nous avons mis en évidence que la description comportementale ne peut se limiter à du gestuel pur, du point de vue modalité, et à l'expression d'états émotionnels du point de vue des fonctions. Les micro-expressions visibles et audibles que nous avons relevées apparaissent hautement informatives et en lien étroit avec les états du *FoT*.

En somme, notre étude a cherché à indiquer des phénomènes et des paramètres pertinents d'un point de vue communicatif, et pourtant souvent négligés par les différentes modélisations, ou générés de manière superficielle sans tenir compte de la globalité de l'information apportée. Or, cette étude approfondie du comportement de deux et six sujets de culture française (respectivement pour les modalités visuelles et acoustiques) nous permet d'insister sur la cohérence indispensable du comportement à modéliser.

Afin qu'une interaction apparaisse réaliste et écologique, il est donc nécessaire de modéliser dans les technologies la variation du comportement dans sa multimodalité. Cette modélisation doit de plus s'attacher à être cohérente en soi et personnifiée, mais aussi cohérente par rapport à la dynamique temporelle de la situation de communication. Ainsi, l'interaction doit être considérée dans sa globalité, en impliquant un ou des individus avec leur spécificités, placés une situation particulière.

Lorsque les technologies auront permis de valider une telle modélisation du comportement expressif du *FoT*, nous pourrons alors envisager la modélisation globale de l'expressivité dans une interaction entre deux humains, et au sein d'une communication située.

# BIBLIOGRAPHIE

(2009). Lévi-Strauss le dernier géant. *Marianne, Le Magazine Littéraire*.

Adam, C., & Evrard, F. (2005). Donner des émotions aux agents conversationnels. In *Workshop Francophone sur les Agents Conversationnels Animés (WACA'01)* (pp. 135-144). Grenoble.

Adda-Decker, M. (2006). De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. In *XXVIème Journées d'Etude sur la Parole (JEP)* (pp. 389-400). Dinard.

Adell, J., Bonafonte, A., & Escudero-Mancebo, D. (2010). Modelling filled pauses prosody to synthesise disfluent speech. In *Speech Prosody*. Chicago, USA.

Alwood, J. (1995). an activity based approach to pragmatics (GPTL) 75, Papers in Theoretical Linguistics. Gothenburg.

Ameka, F. (1992). Interjections: The universal yet neglected part of speech. *Journal of Pragmatics*, *18*, 101-118.

André, E., Dybkjær, L., Minker, W., & Heisterkamp, P. (2004). Affective Dialogue Systems : Tutorial and Research Workshop, ADS 2004. Kloster Irsee, Germany: Springer, Berlin.

Ardissono, L., Boella, G., & Lesmo, L. (1999). Politeness and Speech Acts. In *First Workshop on 'Attitudes, Personality and Emotions in User-Adapted Interactions'* (pp. 41-55).

Arnold, M. (1960). *Emotion and personality (vol. 1 & 2). New York : Columbia University Press. Aubergé, V*. New York: Columbia University Press.

Aubergé, V. (2002). A Gestalt morphology of prosody directed by functions : the example of a step by step model developed at ICP. In *Proceedings of 1st International Conference on Speech Prosody* (pp. 151-155). Aix-en-Provence.

Aubergé, V. (2002). Prosodie et émotion. In *Actes des deuxièmes assises nationales du GdR I3* (pp. 263-273). Nancy.

Aubergé, V., & Cathiard, M. (2003). Can we hear the prosody of smile? *Speech Communication*, *40*(1-2), 87-97.

Aubergé, V., Audibert, N., & Rilliard, A. (2006). De E-Wiz à C-Clone. Recueil, modélisation et synthèse d'expressions authentiques. *Revue d'Intelligence Artificielle - "Interactions émotionnelles"*, *20*(4-5), 499-528.

Audibert, N. (2004). E-Wiz: capture, analyse perceptive et acoustique de parole expressive. Mémoire de Master. Université Grenoble 3.

Audibert, N. (2008). Prosodie de la parole expressive : dimensionnalité d'énoncés. Ph.D. Thesis. Institut National Polytechnique de Grenoble.

Audibert, N., Aubergé, V., & Rilliard, A. (2007). When is the Emotional Information? A gating experiment for gradient and contours cues. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)* (pp. 2137-2140). Saarbrücken, Allemagne.

Austin, J. L. (1975). *How to Do Things with Words: Second Edition (William James Lectures)*. Harvard University Press.

Averill, J. (1975). A Semantic Atlas of Emotional Concepts. *JSAS Catalog of Selected Documents in Psychology*, *5*(330).

Averill, J. (1980). A constructivist view of emotion. In R. Plutchik & H. Kellerman, *Theories of emotion* (pp. 305-339). New York: Academic Press.

Bachmann, C., Lindenfeld, J., & Simonin, J. (1981). *Langage et communications sociales*. Paris: Hatier-Crédif.

Barclay, C. D., Cutting, J. E., & Kozlowski, L. T. (1978). Temporal and spatial factors in gait perception that influence gender recognition. *Perception & psychophysics*, *23*(2), 145-52.

Barkhuysen, P., Krahmer, E., & Swerts, M. (2005). Problem detection in human?machine interactions based on facial expressions of users. *Speech Communication*, *45*(3), 343-359.

Barkhuysen, P., Krahmer, E., & Swerts, M. (2010). Cross-modal and incremental perception of audiovisual cues to emotional speech. *Language and Speech*, *In press*.

Barnouw, V. (1985). *Culture and personality*. Chicago, USA: Dorsey Press.

Bassili, J. N. (1978). Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(3), 373-379.

Bassili, J. N. (1979). Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of personality and social psychology*, *37*(11), 2049-58.

Baudouin, J. (2001). Reconnaissance du visage, expression et genre. Ph.D. Thesis. Université Lyon 2.

Benoît, C., & Le Goff, B. (1998). Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP. *Speech Communication*, *26*(1-2), 117-129.

Berlin, B. (1994). Evidence for pervasive synesthetic sound symbolism in ethnozoological nomenclature. In L. Hinton, J. Nichols, & J. Ohala, *Sound symbolism* (pp. 525-568). Cambridge University Press.

Berthold, A., & Jameson, A. (1999). Interpreting symptoms of cognitive load in speech input. In J. Kay, *Proceedings of User Modeling '99* (pp. 235-244).

Beskow, J. (1995). Rule-Based Visual Speech Synthesis. In *Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH '95)* (pp. 299-302). Madrid, Spain.

Beun, R., De Vos, E., & Witteman, C. (2003). Embodied Conversational Agents: Effects on memory performance and anthropomorphisation. *Lecture notes in computer science, proceeding of IVA'03*, *2792*, 315-319. Kloster Irsee, Germany: Springer.

Bevacqua, E., Heylen, D. K., Pelachaud, C., & Tellier, M. (2007). Facial Feedback Signals for ECAs. In *AISB'07: Artificial and Ambiant Intelligence*. Newcastle University, Newcastle upon Tyne, UK.

Biederman, I. (1987). Recognition-by-components: A theory of human image interpretation. *Psychological Review*, *94*, 115-148.

Bindemann, M., Burton, A. M., Langton, S. R., Schweinberger, S. R., & Doherty, M. J. (2007). The control of attention to faces. *Journal of vision*, *7*(10), 15.1-8.

Birdwhistell, R. L. (1968). L'analyse kinésique. (M. Lacoste) *Langages*, *3*(10), 101-106. Persée - Portail des revues scientifiques en SHS.

Blache, P., Bertrand, R., Bigi, B., Bruno, E., Cela, E., Espesser, R., et al. (2010). Multimodal Annotation of Conversational Data. In *Proceedings of Linguistic Annotation Workshop*. Uppsala, Sweden.

Bodamer, J. (1947). Die Prosop-Agnosie. *Archiv für Psychiatrie und Nervenkrankheiten Vereinigt mit Zeitschrift für die Gesamte Neurologie und Psychiatrie*, *179*(1-2), 6-53.

Borod, J. C., Pick, L. H., Hall, S., Sliwinski, M., Madigan, N., Obler, L. K., et al. (2000). Relationships among facial, prosodic, and lexical channels of emotional perceptual processing. *Cognition and Emotion*, *14*(2), 193-211.

Boyer, J. (1997). Effets de la simultanéité de production entre gestes iconiques ou métaphoriques et contenus verbaux. *Travaux de l'Institut de phonétique d'Aix*, *17*, 249-266. Institut de phonétique.

Brennan, S. E., & Williams, M. (1995). The Feeling of Another's Knowing: Prosody and filled pauses as cues to listeners about metacognitive states of speakers. *Journal of memory and language*, *34*(3), 383-398.

Bruce, V., & Valentine, T. (1988). When a nod's as good as a wink: The role of dynamic information in facial recognition. In M. Gruneberg, P. Morris, & R. Sykes, *Practical aspects of memory: Current research and Issues* (Vol.1.). Chichester: John Wiley.

Buck, R. (1985). Prime theory : an integrated view of motivation and emotion. *Psychological review*, *92*(3), 389-413. American Psychological Association.

Buisine, S., Abrilian, S., Niewiadomski, R., Martin, J. C., Devillers, L., Pelachaud, C., et al. (2006). Perception of blended emotions: From video corpus to expressive agent. In *Intelligent Virtual Agents, Proceedings* (Vol. 4133, pp. 93-106).

Buisine, S., Hartmann, B., Mancini, M., & Pelachaud, C. (2006). Conception et évaluation d'un modèle d'expressivité pour les gestes des agents conversationnels. *Revue d'intelligence artificielle*, *20*(4-5), 621-638. Lavoisier.

Bänziger, T., & Scherer, K. (2007). Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus. In A. C. Paiva, R. Prada, & R. W. Picard, *Affective Computing and Intelligent Interaction*, Lecture Notes in Computer Science (Vol. 4738, pp. 476-487). Berlin, Heidelberg: Springer.

Böhm, T., & Shattuck-Hufnagel, S. (2009). Do Listeners Store in Memory a Speaker's Habitual Utterance-Final PhonationType? *Phonetica*, *66*(3), 150-168. Karger.

Calbris, G. (2003). *L'expression gestuelle de la pensée d'un homme politique*. Paris: CNRS Edition.

Calbris, G., & Porcher, L. (1989). *Geste et communication*. Paris: Credif-Hatier.

Calder, A. J., Young, A. W., Keane, J., & Dean, M. (2000). Configural information in facial expression perception. *Journal of experimental psychology. Human perception and performance*, *26*(2), 527-551. American Psychological Association.

Campbell, N. (2000). Databases of emotional speech. In *ITRW on Speech and Emotion* (pp. 34-38). Newcastle, Northern Ireland, UK.

Campbell, N. (2004). Getting to the Heart of the Matter: Speech as the Expression of Affect; Rather than Just Text or Language. In *Languages Resources and Evaluation* (Vol. 39, pp. 109-118).

Campbell, N. (2007). On the Use of NonVerbal Speech Sounds in Human Communication. In A. Esposito, M. Faundez-Zanuy, E. Keller, & M. Marinaro, *Verbal and Nonverbal Communication Behaviours*, Lecture Notes in Computer Science (Vol. 4775, pp. 117-128). Berlin, Heidelberg: Springer.

Campbell, N., Devillers, L., Douglas-cowie, E., Aubergé, V., Batliner, A., Tao, J., et al. (2006). Resources for the Processing of Affect in Interactions. In *Panel session of LREC'06* (pp. xxiv-xxvii).

Campbell, N., Kashioka, H., & Ohara, R. (2005). No laughing matter. In *Proceedings of INTERSPEECH 2005* (p. 465–468).

Campbell, R., Landis, T., & Regard, M. (1986). Face recognition and lipreading: A neurological dissociation. *Brain*, *109*(3), 509-521.

Candea, M., Vasilescu, I., & Adda-Decker, M. (2005). Inter- and intra-language acoustic analysis of autonomous fillers. In *DISS 05, Disfluency in Spontaneous Speech Workshop*. Aix-en-Provence, France.

Cannon, W. B. (1927). The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory. *The American Journal of Psychology*, *39*(1), 106-124.

Carberry, S., & Schroeder, L. (2001). Recognizing and conveying attitude and its underlying motivation. In *Second Workshop on 'Attitudes, Personality and Emotions in User-Adapted Interaction'*.

Carlier, P., & Graff, C. (2006). Unpredictability as a counter strategy : An analysis of elite tennis matches. *Journal of Sports Sciences*.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (2000). *Embodied Conversational Agents*. Cambridge, MA: The MIT Press.

Cassell, J., Vilhjálmsson, H. H., & Bickmore, T. (2001). BEAT: the Behavior Expression Animation Toolkit. In *International Conference on Computer Graphics and Interactive Techniques* (pp. 477-486).

Christophe, V. (1998). *Les émotions: tour d'horizon des principales théories*. Villeuneuve d'Ascq: Presses Universitaires du Septentrion.

Clark, H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*(1), 62-81.

Cohn, J. F. (2007). Foundations of human computing: Facial expression and emotion. In *State of the Art Survey* (Vol. 4451, pp. 1-16). Berlin Heidelberg: Springer.

Cohn, J. F., & Ekman, P. (2005). Measuring facial action. In *The new handbook of Methods in Nonverbal Behavior Research* (pp. 9-64).

Colletta, J., Kunene, R., Venouil, A., Kaufmann, V., & Simon, J. (2009). Multi-track Annotation of Child Language and Gestures. In M. Kipp, J. Martin, P. Paggio, & D. Heylen, *Multimodal Corpora*, Lecture Notes in Computer Science (Vol. 5509, pp. 54-72). Berlin, Heidelberg: Springer.

Contini, M. (1989). L'interjection en Sarde. Une approche linguistique. In *Espaces Romans : Études de dialectologie et de géolinguistique offertes à Gaston Tuaillon* (pp. 320-329). Grenoble: ELLUG.

Cornelius, R. (1996). *The Science of Emotion. Research and Tradition in the Psychology of Emotion*. Upper Saddle River, NJ: Prentice-Hall.

Cornelius, R. (2000). Theoretical Approaches to Emotion. In *ISCA Workshop on Speech and Emotions* (pp. 3-10). Newcastle, Irlande du Nord.

Corraze, J. (1996). *Les communications non-verbales* (5ème éd.). Paris: PUF.

Cosnier, J. (1977). Communication non verbale et langage. *Psychologie Medicale (Numéro spécial Ethologie humaine).*, *9*(11), 2033-2047.

Cosnier, J. (1994). *Psychologie des émotions et des sentiments* (Retz-Natha.). Paris.

Cosnier, J., & Bourgain, D. (1993). Introduction. In R. Pléty, *Ethologie des communication humaine* (ARCI., pp. 7-18). Lyon: Presses Universitaires de Lyon.

Cosnier, J., & Brossard, A. (1984). *La communication non verbale*. *Textes de base*. Neuchâtel: Delachaux & Niestlé.

Courgeon, M., Martin, J. C., & Jacquemin, C. (2008). MARC : un personnage virtuel réactif expressif. In *WACA 2008* (pp. CD-ROM Proceedings). Paris.

Cowie, R. (2005). Emotion-oriented computing: State of the art and key challenges. Report of the Humaine Network of Excellence.

Cowie, R. (2009). Perceiving emotion: towards a realistic understanding of the task. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *364*(1535), 3515-25.

Cruz, M. P. (2009). Towards an Alternative Relevance-Theoretic Approach to Interjections. *International Review of Pragmatics*, *1*(1), 182-206. BRILL.

Cutting, J., & Kozlowski, L. (1977). Recognizing friends by their walk : Gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, *9*, 353-356.

Damasio, A. (1994). *Descartes'error. Emotion, reason, and the human brain* (Putnam Boo.).

Darwin, C. (1872). *The expression of the emotions in man and animals* (Vol. 6). New York, USA: Philosophical Library.

Davidson, R. J., Scherer, K. R., & Goldsmith, H. H. (2003). *Handbook of Affective Sciences*. *Series in Affective Sciences*. Oxford University Press.

De Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion*, *14*(3), 289-311. Psychology Press.

De Rosis, F. (2001). Preface: Towards Adaptation of Interaction to Affective Factors. In *User Modeling and User-Adapted Interaction* (Vol. 11, pp. 267-278). Kluwer Academic Publishers.

De Rosis, F., Carofiglio, V., Grassano, G., & Castelfranchi, C. (2003). Can Computers Deliberately Deceive? A Simulation Tool and Its Application to Turing's Imitation Game. *Computational Intelligence*, *19*(3), 235-263. Blackwell Publishing, Inc.

De Sevin, E., Hyniewska, S. J., & Pelachaud, C. (2010). Influence des Traits de Personnalité sur la Sélection des Rétroactions. In *WACA 2010*. Lille.

Deacon, T. (1997). *The Symbolic Species: The Co-evolution of Language and the Brain*. New York: W.W Norton.

Desimone, R., & Ungerleider, L. (1989). Neural mechanisms of visual processing in monkeys. In F. Boller & J. Grafman, *Handbook of Neuropsychology* (Vol.2.). Amsterdam: Elsevier Science.

Dijkstra, C., Krahmer, E., & Swerts, M. (2006). Manipulating Uncertainty: The Contribution of Different Audiovisual Prosodic Cues to the Perception of Confidence. In *Speech Prosody 2006* (p. 025). Dresden, Germany.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, *40*(1-2), 33-60.

Ducrot, O., & Schaeffer, J. (1995). *Nouveau dictionnaire encyclopédique des sciences du langage*. Editions du seuil.

Dumas, G. (1948). *La vie affective*. Paris: Presses Universitaires de France.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, *23*(2), 283-292.

Dymond, R. F. (1949). A scale for measurement of empathic ability. *Journal of Consulting and Clinical Psychology*, *13*(2), 127-133.

Eastman, C. M. (1992). Swahili interjections: Blurring language-use/gesture-use boundaries. *Journal of Pragmatics*, *18*(2-3), 273-287.

Egges, A., Kshirsagar, S., & Magnenat-Thalmann, N. (2004). Generic personality and emotion simulation for conversational agents. *Computer Animation and Virtual Worlds*, *15*(1), 1-13.

Eklund, R. (2008). Pulmonic ingressive phonation : Diachronic and synchronic characteristics, distribution and function in animal and human sound production and in human speech. *Journal of the International Phonetic Association*, *38*(3), 235-324. Cambridge University Press.

Ekman, P. (1989). L'Expression des Emotions. In B. Rimé & K. R. Scherer, *Les Emotions* (pp. 183-201). Neuchâtel ; Paris: Delachaux-Niestlé.

Ekman, P. (1999). Facial expressions. In *Handbook of Cognition and Emotion* (pp. 301-320). John Wiley & sons Ltd.

Ekman, P. (2003). *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Times books.

Ekman, P., & Friesen, W. (1969). The Repertoire Of Nonverbal Behavior: Categories, Origins, Usage and Coding. *Semiotica*, *1*, 49-98.

Ekman, P., & Friesen, W. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124-129.

Ekman, P., & Friesen, W. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, California: Consulting Psychologists Press.

Ekman, P., & Friesen, W. V. (1975). *Unmasking the face. A guide to recognizing emotions from facial clues*. Prentice Hall Trade.

Ekman, P., & Friesen, W. V. (1976). Measuring Facial Movement. *The Journal of Environmental Psychology and Nonverbal Behavior*, *1*(1), 56-75. New York: Human Sciences Press.

Ekman, P., & Oster, H. (1979). Facial Expressions of Emotion. *Annual Review of Psychology*, *30*, 527-554.

Ekman, P., Friesen, W. V., & Tomkins, S. (1971). Facial Affect Scoring Technique: A First Validity Study. *Semiotica*, *3*(1), 37-58. De Gruyter.

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science*, *221*, 1208-1210.

Elliott, C. (1999). Why boys like motorcycles: using emotion theory to find structure in humorous stories. In *Proceedings of the Workshop on EBAA '99*.

Elliott, R. (1998). A Model Of Emotion-Driven Choice. *Journal of Marketing Management*, *14*(1), 95-108. Routledge, part of the Taylor & Francis Group.

Ellison, J. W., & Massaro, D. W. (1997). Featural Evaluation, Integration, and Judgment of Facial Affect. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(1), 13.

Fauve, B., Matrouf, D., Scheffer, N., Bonastre, J., & Mason, J. S. (2007). State-of-the-Art Performance in Text-Independent Speaker Verification Through Open-Source Software. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(7), 1960-1968.

Feyereisen, P. (1994). *Le cerveau et la communication*. Paris: Presses Universitaires de France.

Fodor, J. (1983). *Modularity of mind*. Cambridge, MA: MIT Press.

Foulquié, P. (1962). *Dictionnaire de la langue philosophique*. Paris: PUF.

Frasson, C., & Gauthier, G. (1990). *Intelligent Tutoring Systems: At the Crossroads of Artificial Intelligence and Education*. Westport, CT, USA: Greenwood Publishing Group Inc.

Fredouille, C. (2002). Reconnaissance Automatique du Locuteur, performances et Limites. In *6ème Congrès d'Acoustique – Session spéciale: Les expertises vocales: l'identification en question*. Lille.

Fredrickson, B., & Levenson, R. W. (1998). Positive emotion speed recovery from the cardiovascular sequelae of negative emotions. *Cognition & Emotion*, *12*, 191-220.

Fridlund, A. J. (1995). *Human Facial Expression: An Evolutionary View* (p. 369). Academic Press Inc.

Frijda, N. H. (1986). *The emotions*. Cambridge and New York: Cambridge University Press.

Frijda, N. H. (1987). Emotion, cognitive structure, and action tendency. *Cognition & Emotion*, *1*(2), 115-143. Psychology Press.

Frijda, N. H., Kuipers, P., & ter Schure, E. (1989). Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, *57*(2), 212-228.

Fónagy, I. (1991). *La vive voix. Essais de psycho-phonétique* (2nd éd., p. 346). Paris: Payot.

Garfinkel, H. (1967). *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.

Gaver, W. (2009). Designing for emotion (among other things). *Philosophical Transactions of the Royal Society, Biological sciences*, *364*, 3597-3604.

Gebhard, P., Klesen, M., & Rist, T. (2004). Coloring Multi-character Conversations through the Expression of Emotions . In E. André, L. Dybkjær, W. Minker, & P. Heisterkamp, *Proceeding of the Tutorial and Research Workshop on Affective Dialogue Systems (ADS'04)*, Lecture Notes in Computer Science (Vol. 3068, pp. 128-141). Kloster Irsee: Springer Berlin Heidelberg.

Gergen, K. J. (1985). The social constructionist movement in modern psychology. *The American psychologist*, *40*(3), 266-275. American Psychological Association.

Gibet, S., Kamp, J., & Poirier, F. (2004). Gesture Analysis: Invariant Laws in Movement. In A. Camurri & G. Volpe, *Gesture-Based Communication in Human-Computer Interaction,* Lecture Notes in Computer Science (GW 2003 :., Vol. 2915, pp. 1-9). Springer Berlin / Heidelberg.

Goffman, E. (1963). *Behaviour in public places*. Illinois: Free Press.

Goffman, E. (1981). *Forms of Talk*. Oxford: Oxford University Press.

Goffman, E. (1981). Response cries. In *Forms of Talk* (pp. 78-122). Philadelphia: University of Pennsylvania Press.

Goldie, P., Döring, S., & Al. (2005). Humaine deliverable D10d: Interim Report to plenary meeting on ethical frameworks for emotion-oriented systems.

Goodwin, C. (1981). *Conversational Organization - interaction between speakers and hearers*. Academic Press.

Goto, M., Itou, K., & Hayamizu, S. (1999). A real-time filled pause detection system for spontaneous speech recognition. In *Eurospeech 1999* (p. 227–230).

Graff, C. (2008). Notes de cours d'"éthologie et chronobiologie." Université Pierre Mendès France – Grenoble.

Graff, C. (2009). Piste méthodologiques et conceptuels de la biologie du comportements applicable à l'IHM, Séminaire au LIG. Grenoble.

Grammer, K., & Oberzaucher, E. (2006). The Reconstruction of Facial Expressions in Embodied Systems: New Approaches to an Old Problem. *ZiF: Mitteilungen*, *2*.

Grandjean, D., & Baenzinger, T. (2009). Expression Vocale Des Emotions. In D. Sander & K. Scherer, *Traité de psychologie des émotions*. Paris: Dunod.

Grandjean, D., & Scherer, K. R. (2009). Théorie de l'évaluation cognitive et dynamique des processus émotionnels. In D. Sander & K. R. Scherer, *Traité de psychologie des émotions* (pp. 42-76). Paris: Dunod.

Grice, H. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Guaïtella, I., Santi, S., Cavé, C., Bertrand, R., Boyer, J., Faraco, M., et al. (1998). Les relations voco-gestuelles dans la communication interpersonnelle. In *ORAGE'98, ORAlité et Gestualité: communication multimodale, interaction* (pp. 13-24). Paris: L'Harmattan.

Gussenhoven, C. (2002). Intonation and Interpretation: Phonetics and Phonology. In *Speech Prosody 2002* (pp. 47-57). Aix-en-Provence, France.

Hager, J. C. (1983). The inner and outer meanings of facial expressions. In J. Cacioppo & J. Hager, *Social Psychphysiology: A Sourcebook*. New York: Guilford.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of educational psychology*, *56*(4), 208-16.

Hauser, M. D. (1997). *The Evolution of Communication*. Cambridge: The MIT Press.

Heisterkamp, P. (2003). "Do not attempt to light with match!": some thoughts on progress and research goals in spoken dialog systems. In *Eurospeech*.

Hess, U., & Kirouac, G. (2003). Emotion expression in groupsNo Title. In M. Lewis & J. M. Haviland-Jones, *Handbook Of Emotions*. New-York, USA: The Guilford Press.

Heylen, D. K. (2007). Multimodal Backchannel Generation for Conversational Agents. In *Workshop on Multimodal Output Generation (MOG 2007)*. Aberdeen, Scotland.

Heylen, D. K., Nijholt, A., & Poel, M. (2007). Generating Nonverbal Signals for a Sensitive Artificial Listener. In *Verbal and Nonverbal Communication Behaviours* (pp. 264-274). Heidelberg: Springer Berlin.

Humphreys, G., Donnelly, N., & Riddoch, M. (1993). Expression is computed separately from facial identity, and it is computed separately for moving and static faces : neuropsychological evidence. *Neuropsychologia*, *31*(2), 173-181. Elsevier.

Hénaff, M. (2009). Lévi-Strauss le dernier géant, "La nouveauté structurale." *Marianne, Le Magazine Littéraire, Hors-série Nov-Déc. 2009*.

Ickes, W. (2009). Empathic Accuracy: Its Links to Clinical, Cognitive, Developmental, Social, and Physiological Psychology. In J. Decety & W. Ickes, *The Social Neuroscience of Empathy* (pp. 57-70). Cambridge, MA: MIT Press.

Izard, C., & Ackerman, B. (2000). Motivational, organizational, and regulatory functions of discrete emotions. In M. Lewis & Haviland-Jones, *Handbook of emotions* (Vol. 2, p. 253–264). Guilford Press.

Izaute, M., Chambres, P., & Larochelle, S. (2002). Feeling-of-knowing for proper names. *Canadian Journal of Experimental Psychology*.

Jackendoff, R. (1999). Possible stages in the evolution of the language capacity. *Trends in Cognitive Sciences*, *3*(7), 272-279.

Jackendoff, R. (2002). An evolutionary perspective on the architecture. In *Foundations of Language: Brain, Meaning, Grammar, Evolution* (pp. 231-264). Oxford: Oxford University Press.

James, W. (1884). What Is An Emotion? *Mind*, *9*, 188-205.

James, W. (1890). *The principles of Psychology*. Dover Publications.

James, W. (1892). *Psychology: Briefer Course*. New York: Henry Holt.

Janet, P. (1926). La pensée intérieure et ses troubles: Compte rendu intégral du cours professé par P. Janet au Collège de France. In *Psychologie expérimentale et comparée*. Paris: Publications A. Chahine.

Jespersen, O. (1921). *Language: Its Nature, Development, and Origin*. London: Allen and Unwin.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, *14*, 201-211.

Johnstone, T., van Reekum, C. M., Hird, K., Kirsner, K., & Scherer, K. R. (2005). Affective Speech Elicited With a Computer Game. *Emotion*, *5*(4), 513-518. Tom Johnstone, US: American Psychological Association.

Juslin, P. N., & Scherer, K. R. (2005). Vocal expression of affect. In *The new handbook of Methods in Nonverbal Behavior Research* (pp. 65-135). Oxford University Press.

Kainz, F. (1962). *Psychologie der Sprache: Grundlagen der allgemeinen Sprachpsychologie* (3rd ed.). Stuttgart: Enke.

Kaiser, S., & Wherle, T. (2006). Modeling appraisal theory of emotion and facial expression. In *19th International Conference on Computer Animation and SocialAgents, CASA 2006*. Genève.

Kaiser, S., Wehrle, T., & Schenkel, K. (2009). Expression faciale. In *Traité de psychologie des émotions* (pp. 77-108). Paris: Dunod.

Kaiser, S., Wehrle, T., & Schmidt, S. (1998). Emotional episodes, facial expressions, and reported feelings in human-computer interactions. In A. Fischer, *Proceedings of the Xth Conference of the International Society for Research on Emotions* (pp. 82-86). Würzburg: ISRE Publications.

Kant, E. (1781). *Critique de la raison pure*. (J. Hartknoch) (Vol. 1). Riga.

Kendon, A. (1988). How gestures can become like words. In F. Poyatos, *Cross-Cultural Perspectives in Nonverbal Communication* (pp. 131-141). Lewiston, New York: C.J. Hogrefe.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge University Press.

Kerbrat-Orecchioni, C. (1986). « Nouvelle communication » et « analyse conversationnelle ». *Langue française*, *70*(1), 7-25.

Kleinpaul, R. (1888). *Sprache ohne Worte. Idee einer allgemeinen Wissenschaft der Sprache*. Leipzig (New printing The Hague: Mouton, 1972).

Kompe, R. (1997). Prosody in speech understanding systems. *Lecture notes in computer science*, *1307*, xix, 357 p. Springer.

Krahmer, E., & Swerts, M. (2009). Audiovisual prosody - Introduction to the Spécial Issue. *Language and Speech*, *52*(2), 129-133.

Kraut, R. E., & Johnston, R. E. (1979). Social and emotional messages of smiling: An ethological approach. *Journal of Personality and Social Psychology*, *37*(9), 1539-1553.

Labov, W. (1973). Some principles of linguistic methodology. *Language in Society*, *1*, 97-120.

Lacheret-Dujour, A., & Beaugendre, F. (1999). *La Prosodie du français*. Paris: Edition du CNRS.

Lange, C. (1885). *Om Sindsbevaegelser et Psyko-Fysiologisk Studie*. Copenhague: Rasmussen.

Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of speech, language, and hearing research : JSLHR*, *42*(3), 526-39.

Lanzetta, J. T., Cartwright-Smith, J., & Eleck, R. E. (1976). Effects of nonverbal dissimulation on emotional experience and autonomic arousal. *Journal of Personality and Social Psychology*, *33*(3), 354-370.

Lardellier, P. (2008). *Arrêtez de décoder*. Charmey, Suisse: Les édition de l'Hèbe.

Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press.

Le Moigne, J. (1995). *Les épistémologies constructivistes*. Paris: Presses Universitaires de France, "Que sais-je?."

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., Bhogal, R. S., et al. (1997). The persona effect: affective impact of animated pedagogical agents. *Conference on Human Factors in Computing Systems*, 359-366. Atlanta, Georgia, USA.

Lester, J. C., Towns, S. G., Callaway, C. B., Voerman, J. L., & FitzGerald, P. J. (2000). Deictic and emotive communication in animated pedagogical agents. In J. Cassell, S. Prevost, & J. Sullivant, *Embodied conversational agents* (pp. 123-154). Boston: MIT Press.

Li, C. (2005). A Cognitive-pragmatic Account of Interjections. *US-China Foreign Language*, *3*(9 serial n°24), 65-70.

Lickley, R., Shillcock, R., & Bard, E. (1991). Processing disfluent speech: How and when are disfluencies found? In *Proceedings of European Conference on Speech Technology, vol. 3* (p. 1499–1502).

Lorenz, K. (1965). *Über tierisches und menschliches Verhalten : Aus dem Werdegang der Verhaltenslehre*. München: Pierper.

Loyau, F. (2007). Expressions des états mentaux et émotionnels de l'humain en interaction : ébauches du "Feeling of Thinking". Ph.D. Thesis. Institut National Polytechnique de Grenoble.

Loyau, F., & Aubergé, V. (2006). Expressions outside the talk turn: ethograms of the feeling of thinking. In *5th LREC* (pp. 47-50).

Léon, P. (2005). *Précis de phonostylistique : Parole et expressivité*. Paris: Armand Colin.

Lévi-Strauss, C. (1958). *Anthropologie structurale* (Plon.). Paris.

Lévi-Strauss, C. (1962). *La pensée sauvage* (Plon.). Paris.

Lévi-Strauss, C. (1988). *De près et de loin, entretien avec Didier Eribon* (Odile Jaco.).

Mac, D., Aubergé, V., Rilliard, A., & Castelli, E. (2010). Cross-cultural perception of Vietnamese Audio-Visual prosodic attitudes. In *Speech Prosody 2010*. Chicago, Illinois, USA.

Maclay, H., & Osgood, C. (1959). Hesitation phenomena in spontaneous English speech. *Word*, *15*, 19-44.

Magnenat-Thalmann, N., HyungSeok, K., Egges, A., & Garchery, S. (2005). Believability and Interaction in Virtual Worlds. In *11th International Multimedia Modelling Conference (MMM'05)* (pp. 2-9). Melbourne, Australia.

Magno Caldognetto, E., & Poggi, I. (1997). Micro- and Macro-Bimodality. In R. Campbell & C. Benoit, *Proceedings of the Workshop on Audio Visual Speech Perception* (pp. 33-36). Rhodes, Greece.

Malrieu, P. (1952). *Les émotions et la personnalité de l'enfant*. Paris: Vrin, 2nd éd.

Mancini, M. (2008). Agents conversationnels animés avec comportements distinctifs. Ph.D. Thesis. Université de Paris 8.

Markus, H., & Kitayama, S. (1998). The Cultural Psychology of Personality. *Journal of Cross-Cultural Psychology*, *29*(1), 63-87.

Marr, D. (1983). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt & Company.

Martin, J., Abrilian, S., Devillers, L., Lamolle, M., Mancini, M., Pelachaud, C., et al. (2006). Du corpus vidéo à l'agent expressif : Utilisation des différents niveaux de représentation multimodale et émotionnelle. In *Revue des sciences et technologies de l'information. Série Revue d'intelligence artificielle* (Vol. 20, pp. 477-498).

Martin, J., Niewiadomski, R., Devillers, L., Buisine, S., & Pelachaud, C. (2006). Multimodal complex emotions: Gesture expressivity and blended facial expressions. *International Journal of Humanoid Robotics (IJHR)*, *3*(3), 269-291.

Martin, P. (2004). WinPitch LTL, un logiciel d'enseignement de la prosodie multimédia. In *TALAL* (pp. 71-82). Grenoble, France.

Martin, P. (2006). Intonation du français: parole spontanée et parole lue. *EFE*, *XV*, 135-162.

Martin, P. (2008). *Phonétique acoustique : Introduction à l'analyse acoustique de la parole* (Collection.). Paris: Armand Collin.

Martin, P. (2011). La prosodie du français: une approche pas très syntaxique. *Journal of French Language Studies*, *21*, 39-52.

Martinet, A. (1964). *Elements of General Linguistics*. University of Chicago.

Massaro, D., & Egan, P. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin and Review*, *3*, 215-221.

Matsumoto, D. (2006). Are cultural differences in emotion regulation mediated by personality traits? *Journal of Cross-Cultural Psychology*, *37*(4), 421-437.

Matsumoto, D. (2007). Culture, context, and behavior. *Journal of personality*, *75*(6), 1285-319.

Matsumoto, D., Yoo Hee, S., & Fontaine, J. (2008). Mapping Expressive Differences Around the World: The Relationship Between Emotional Display Rules and Individualism Versus Collectivism. *Journal of Cross-Cultural Psychology*, *39*(1), 55-74.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.

McNeill, D. (1992). *Hand and mind. What gestures reveal about thought*. The university of Chicago Press.

McNeill, D. (2005). *Gesture and Thought*. University Of Chicago Press.

Mehrabian, A., & Wiener, M. (1967). Decoding of inconsistent communications. *Journal of Personality and Social Psychology*, *6*, 109-114.

Merleau-Ponty, M. (1976). *Phénoménologie de la perception*. Paris: Edition Gallimard, collection "Tel."

Merleau-Ponty, M. (1990). *La structure du comportement*. Paris: Presses Universitaires de France, collection "Quadrige."

Messing, L. (1996). What's the Use of Bimodal Communication? In L. Messing, *Proceedings of the WIGLS, "Workshop on the Integration of Gesture and Language in Speech* (pp. 115-124). Newark and Wilmington, Delaware.

Minsky, M. (1986). *The society of mind*. New York, NY, USA: Simon & Schuster, Inc.

Molénat, X. (2008). Ethnométhodologie, la société en pratiques. *Sciences humaines, rubrique "Références"*.

Mondada, L. (2008). La transcription dans la perspective de la linguistique interactionnelle. In M. Bilger, *Donnees orales, les enjeux de la transcription* (pp. 78-109). Perpignan: Presses Universitaires de Perpignan.

Mozziconacci, S. J. (1998). *Speech variability and emotion : production and perception*. Technische Universiteit Eindhoven.

Munhall, K., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception. *Psychological Science*, *15*(2), 133-137. SAGE Publications.

Nespolous, J., & Lecours, A. (1986). Gestures: Nature and Function. In J. Nespolous, P. Perron, & A. Lecours, *The Biological Foundations of Gestures: Motor and Semiotic Aspects* (pp. 49-62). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nhouyvanisvong, A., & Reder, L. M. (1998). Rapid Feeling-of-knowing : A Strategy Selection Mechanism. *Metacognition : Cognitive and Social dimensions*, 35-52. London, Sage.

Niedenthal, P. M. (2007). Embodying emotion. *Science*, *316*(5827), 1002-1005.

Nielsen, K. J., Logothetis, N. K., & Rainer, G. (2006). Discrimination Strategies of Humans and Rhesus Monkeys for Complex Visual Displays. *Current Biology*, *16*(8), 814-820.

Nugier, A. (2009). Histoire et grands courants de recherche sur les émotions. *Revue électronique de Psychologie Sociale*, *4*, 8-14.

Ochs, E. (1979). Transcription as theory. In E. Ochs & B. Schieffelin, *Developmental pragmatics* (pp. 43-72). New York: Academic.

Ohala, J. J. (1994). The frequency codes underlies the sound symbolic use of voice pitch. In J. J. Ohala, L. Hinton, & J. Nichols, *Sound symbolism* (pp. 325-347). Cambridge: Cambridge University Press.

Ohala, J. J. (1996). Ethological theory and the expression of emotion in the voice. In *4th International Conference on Spoken Language Processing (ICSLP 96)* (Vol. 3, pp. 1812-1815). Philadelphia, PA, USA: Wilmington : University of Delaware.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge University Press.

Pandzic, I. S., & Forchheimer, R. (2003). *MPEG-4 Facial Animation: The Standard, Implementation and Applications; electronic version*. Chichester: Wiley.

Pantic, M., & Rothkrantz, L. J. (2003). Toward an affect-sensitive multimodal human-computer interaction. In *Proceedings of the Ieee* (Vol. 91, pp. 1370-1390).

Pelachaud, C. (2006). Introduction. *Revue d'Intelligence Artificielle - "Interactions émotionnelles"*, *20*(4-5), 441-445.

Pelachaud, C. (2010). GRETA: Embodied Conversational Agent. Retrieved from http://perso.telecom-paristech.fr/~pelachau/Greta/.

Pelachaud, C., & Poggi, I. (2002). Subtleties of facial expressions in embodied agents. *Journal of Visualization and Computer Animation*, *13*(5), 301-312.

Pelachaud, C., André, E., Bevacqua, E., Burkhardt, F., Camurri, A., Castellano, L., et al. (2005). Humaine, WP6: Proposal for exemplars and work toward them: Emotion in Interaction.

Pelachaud, C., Badler, N. I., & Steedman, M. (1996). Generating Facial Expressions for Speech. *Cognitive Science*, *20*(1), 1-46.

Pelachaud, C., Peters, C., Mancini, M., Bevacqua, E., & Poggi, I. (2005). A model of attention and interest using gaze behaviour. In *IVA'05 International Working Conference on Intelligent Virtual Agents*. Greece.

Petridis, S. (2008). Audiovisual laughter detection based on temporal features. In *Proceedings of the 10th international conference on Multimodal interfaces* (pp. 37-44). Chania, Crete, Greece: ACM.

Philippot, P. (2004). Facteurs cognitifs et réactions corporelles dans le processus émotionnel. In G. Kirouac, *Cognition et émotions* (pp. 37-55). Coimbra: PUL.

Picard, R. (1997). *Affective Computing*. Cambridge: MIT Press.

Plutchik, R. (1970). Emotions, Evolution, and Adaptive Processes. In *Feelings and emotions: the Loyola Symposium*. Academic Press.

Plutchik, R. (1980). *Emotion, a psychoevolutionary synthesis*. New York: Harper Row.

Plutchik, R. (1984). Emotions: A general psychoevolutionary theory. In K. R. Scherer & P. Ekman, *Approaches to emotions* (pp. 197-219). Hillsdale, NJ: Erlbaum.

Pléty, R. (1993). *Ethologie des communications humaines. Aide-mémoire méthodologique*. *Collection Ethologie et Psychologie des communications*. ARCi, Presses Universitaires de Lyon.

Poggi, I. (2002). Mind markers. In N. Trigo, M. Rector, & I. Poggi, *Gestures. Meaning and use*. Oporto, Portugal: University Fernando Pessoa Press.

Poggi, I. (2007). *Mind, hands, face and body. A goal and belief view of multimodal communication*. *Körper - Zeichen - Kultur* (p. 433). Berlin: Weidler.

Poggi, I. (2008). The language of interjections. In *Multimodal Signals: Cognitive and Algorithmic Issues, COST 2102 School* (Vol. 5398/2009, pp. 170-186). Vietri, Italy: Springer.

Poggi, I., Pelachaud, C., & De Carolis, B. (2001). To display or not to display? Towards the architecture of a reflexive agent. In *2nd Workshop on Attitude, Personality and Emotions in User-Adapted Interaction. User Modeling 2001*. Sonthofen (Germany).

Poggi, I., Pelachaud, C., & Magno Caldognetto, E. (2004). Gestural Mind Markers in ECAs. In *Gesture-Based Communication in Human-Computer Interaction* (Heidelberg., pp. 338-349). Springer Berlin.

Poggi, I., Pelachaud, C., De Rosis, F., Carofiglio, V., & De Carolis, B. (2005). GRETA. A Believable Embodied Conversational Agent. In O. Stock & M. Zancanaro, *Multimodal Intelligent Information Presentation* (pp. 3-25). Springer, Kluwer.

Pollick, F. E. (2004). The Features People Use to Recognize Human Movement Style. In *Gesture-Based Communication in Human-Computer Interaction* (Vol. 2915, pp. 467-468). Springer Berlin / Heidelberg.

Pooley, T. (1996). *Chtimi: The urban vernaculars of northern France*. Clevedon: Multilingual Matters Ltd.

Popper, K. (1972). *La connaissance objective (Objective Knowledge: An Evolutionary Approach)*. Paris: Aubier, 1991.

Pradines, M. (1954). Sur les conceptions actuelles de l'émotion. In *La psychologie du XXème siècle*. Paris: Presses Universitaires de France.

Raidt, S. (2008). Gaze and face-to-face communication between a human speaker and an embodied conversational agent. Mutual attention and multimodal deixis. Ph.D. Thesis. Institut National Polytechnique de Grenoble.

Reder, L. M., & Ritter, F. (1992). What determines initial feeling of knowing ? familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 435-451.

Redican, W. (1982). An evolutionary perspective on human facial displays. In P. Ekman, *Emotion in the Human Face. 2nd ed* (pp. 212-280). Elmsford, NY: Pergamon.

Reilly, W. S. (1996). Believable Social and Emotional Agents. Ph.D. Thesis. Carnegie Mellon University, Pittsburgh, PA

Rossi, M. (1999). *L'intonation, le systeme du francais : description et modélisation*. Paris: Orphrys Pub., Collection L'Essentiel.

Russell, J., & Barrett-Feldman, L. (1999). Core affect, prototypical emotional episodes and other things called emotion: dissecting the elephant. *Journal of Personality and Social Psychology*, *76*, 37–63.

Sacks, H. (1992). *Lectures on Conversation, Volumes 1 and 2*. Cambridge: Blackwell.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, *50*(4, part. 1), 696-735.

Sadanobu, T. (2004). A natural history of Japanese pressed voice. *Journal of the Phonetic Society of Japan*, *8*(1), 29-44.

Sajjanhar, U., & Ward, N. G. (2006). Automatic Labeling of Back Channels. Technical Report UTEP-CS-06-26. Retrieved from http://www.cs.utep.edu/vladik/2006/tr06-26.pdf

Sander, D., & Scherer, K. R. (2009). *Traité de psychologie des émotions*. Paris: Dunod.

Sansonnet, J. (2006). Agents Conversationnels Animés : Taxinomie, Problématique, Applications. In *Séminaire ISLE-PRESENCE*. Grenoble.

Sapir, E. (1925). A study of phonetic Symbolism. *Journal of Experimental Psychology*, *12*, 225-239.

Sauter, D. a., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(6), 2408-12.

Scherer, K. (1984). On the nature and function of emotion: A component process approach. In K. Scherer & P. Ekman, *Approaches to emotion* (pp. 293-318). NJ: Erlbaum: Hillsdale.

Scherer, K. (2005). *What are emotions? And how can they be measured? Social science information* (Vol. 44, pp. 695-729). Sage.

Scherer, K. R. (1984). Les émotions: fonctions et composantes. *Cahiers de psychologie cognitive*, *4*(1), 9-39. Association pour la diffusion des recherches en sciences cognitives.

Scherer, K. R. (1985). Vocal Affect Signalling: A Comparative approach. In J. S. Rosenblatt, C. Beer, M. Busnel, & P. J. Slater, *Advances in the Study of Behaviour*, Advances in the Study of Behavior (Vol. 15, pp. 189-344). New York: Academic Press.

Scherer, K. R. (1994). Affect bursts. In *Emotions: Essays on emotion theory* (pp. 161-193). Hillsdale: Lawrence Erlbaum Associates, Inc.

Scherer, K. R. (1999). Appraisal theory. In *Handbook of Cognition and Emotion* (pp. 637-663). Chichester, Wiley.

Scherer, K. R. (2000). *Emotion*. *Introduction to Social Psychology: A European perspective* (3rd., pp. 151-191). Oxford, Blackwell.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1), 227-256. Elsevier.

Scherer, K. R., & Ellgring, H. (2007). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion*, *7*(1), 113-130.

Scherer, K. R., & Sangsue, J. (1996). Le système mental en tant que composant de l'émotion. (G. Kirouac)*Geneva Studies in Emotion and Communication*, *10*(1), 1-13. Coimbra, Portugal.

Scherer, K. R., & Wallbott, H. (1994). Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning. *Journal of Personality and Social Psychology*, *66*(2), 310-328.

Scherer, K. R., & Zei, B. (1989). La voix comme indice affectif. *Revue médicale de la suisse romande*, *109*, 61-66.

Scherer, K. R., Ladd, D. R., & Silverman, K. E. (1984). Vocal cues to speaker affect: Testing two models. *The Journal of the Acoustical Society of America*, *76*(5), 1346-1356. ASA.

Scherer, K. R., Wallbott, H., & Summerfield, A. (1986). *Experiencing emotion*. Cambridge, New York: Cambridge University Press.

Scherer, K., Schorr, A., & Johnstone, T. (2001). *Appraisal processes in emotion: theory, methods, research*. Oxford University Press, USA.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological Review*, *61*(2), 81-88.

Schmidt, S. (1998). Les expressions faciales émotionnelles dans le cadre d'un jeu d'ordinateur : reflet de processus d'évaluation cognitive ou d'émotions de base ?Ph.D. Thesis. Université de Genève.

Schröder, M. (2003). Experimental study of affect bursts. *Speech Communication*, *40*(1-2), 99-116. Elsevier Science Publishers B. V.

Schröder, M., & Cowie, R. (2006). Developing a Consistent View on Emotion-Oriented Computing. In S. Renals & S. Bengio, *Machine Learning for Multimodal Interaction*, Lecture Notes in Computer Science (Vol. 3869, pp. 194-205). Springer Berlin / Heidelberg.

Schröder, M., Cowie, R., Heylen, D. K., Pantic, M., Pelachaud, C., Schuller, B., et al. (2008). Towards responsive Sensitive Artificial Listeners. In *Fourth International Workshop on Human-Computer Conversation*. University of Sheffield.

Schröder, M., Heylen, D. K., & Poggi, I. (2006). Perception of non-verbal emotional listener feedback. In *Speech Prosody 2006* (pp. 43-46). Dresde, Deutschland: TUD press.

Schuller, B., Eyben, F., & Rigoll, G. (2008). Static and Dynamic Modelling for the Recognition of Non-verbal Vocalisations in Conversational Speech. *Lecture Notes In Artificial Intelligence*, *5078*, 99-110.

Schultz, T., & Rogina, I. (1995). Acoustic and Language Modeling of Human and Nonhuman Noises for Human-to-Human Spontaneous Speech Recognition. In *Proceedings of ICASSP-1995, vol.1* (p. 293–296). Detroit, Michigan.

Searle, J. (1979). *Expression and Meaning: Studies in the Theory of Speech Acts*. *essay collection* (p. 187). Cambridge University Press.

Seeley, T. D. (1989). Social foraging in honey bees: how nectar foragers assess their colony's nutritional status. *Behavioral Ecology and Sociobiology*, *24*(3), 181-199. Springer Berlin / Heidelberg.

Septseault, C. (2004). *Les modèles émotionnels informatiques*. Retrieved from http://www.cerv.fr/~septseault/uploads/pages/Enseignements/presentation_decembre_2004.pdf.

Shaver, P., Schwartz, J., Kirsona, D., & O'Connor, C. (1987). Emotion Knowledge: Further Exploration of a Prototype Approach. *Journal of Personality and Social Psychology*, *52*(6), 1061-1086.

Shinamura, A. P., & Squire, L. R. (1986). Memory and metamemory : a study of the feeling-of-knowing phenomenon in amnesic patients. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, *12*(3), 452-460.

Shochi, T., Erickson, D., Rilliard, A., Aubergé, V., & Martin, J. (2008). Recognition of Japanese attitudes in Audio-Visual speech. In *Actes de Speech Prosody 2008*. Campinas, Brésil.

Shochi, T., Erickson, D., Rilliard, A., Aubergé, V., & Martin, J. (2008). Recognition of Japanese attitudes in Audio-Visual speech. In *Speech Prosody 2008*. Campinas, Brésil.

Signorello, R., Aubergé, V., Vanpé, A., Granjon, L., & Audibert, N. (2010). À la recherche d'indices de culture et/ou de langue dans les micro-événements audio-visuels de l'interaction face à face. In *WACA 2010*. Lille.

Simon-Thomas, E. R., Keltner, D. J., Sauter, D., Sinicropi-Yao, L., & Abramson, A. (2009). The voice conveys specific emotions: Evidence from vocal burst displays. *Emotion*, *9*(6), 838-846.

Snowdon, C. (2003). Expression Of Emotion In Nonhumen Animals. In R. Davidson, K. Scherer, & H. Goldsmith, *Handbook of Affectives Sciences* (pp. 457-480). New York: Oxford University Press, USA.

Soussignan, R. (2002). Duchenne Smile, Emotional Experience, and Autonomic Reactivity: A Test of the Facial Feedback Hypothesis. *Emotion*, *2*(1), 22.

Sperber, D., & Wilson, D. (1989). *La pertinence : communication et cognition*. Paris: Minuit.

Stock, O., & WP8-Members. (2005). Proceedings of the Workshop "WP8 Emotion in Communication". ITC-Irst Povo, Trento, Italy.

Streri, A. (2001). Comment l'homme perçoit-il le monde? In A. Weil-Barais, *L'homme cognitif* (pp. 97-210). Paris: Presses Universitaires de France.

Suchman, L. (1987). *Plans an situated actions*. Cambridge, UK: Cambridge University Press.

Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, *53*(1), 81-94.

Swerts, M., & Krahmer, E. (2006). The importance of different facial areas for signalling visual prominence. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006)* (p. paper 1289). Pittsburgh, PA, USA: ISCA.

Swerts, M., & Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics*, *36*(2), 219-238.

Tannen, D., & Saville-Troike, M. (1985). *Perspectives on Silence*. Norwood, NJ: Ablex.

Thórisson, K. R. (2002). Natural Turn-Taking Needs No Manual: Computational Theory And Model, From Perception To Action. In B. Granström, D. House, & I. Karlsson, *Multimodality in language and speech systems* (pp. 173-207). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Tinbergen, N. (1963). On aims and methods of Ethology. *Zeitschrift für Tierpsychologie*, *20*(4), 410-433.

Tolkmitt, F. J., & Scherer, K. R. (1986). Effect of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology: Human Perception and Performance*, *12*(3), 302-313.

Tomkins, S. (1980). Affect as amplification: some modifications in theory, NY : Academic Press,. In R. Plutchick & H. Kellerman, *Theory, research and experience: theories of emotions* (pp. 141-165). New York: Academic Press.

Tomkins, S. (1984). Affect theory. In K. R. Scherer & P. Ekman, *Approaches to emotion* (pp. 163-195). Hillsdale, NJ: Erlbaum.

Trager, G. (1958). Paralanguage: A First Approximation. In *Studies in Linguistics* (pp. 1-12).

Treisman, A. (1988). Features and objects: the fourteenth Bartlett memorial lecture. *The Quarterly journal of experimental psychology. A. Human experimental psychology*, *40*(2), 201-237. Taylor & Francis.

Urbain, J., Bevacqua, E., Dutoit, T., Moinet, A., Niewiadomski, R., Pelachaud, C., et al. (2010). La base de données AVLaughterCycle. In *XXVIIIèmes Journées d'Etude sur la Parole (JEP)* (pp. 61-64). Mons, Belgique.

Van Hooff, J. (1972). A comparative approach to the phylogeny of laughter and smiling. In R. Hinde, *Non-verbal communication* (pp. 209-237). Cambridge, England: Cambridge University Press.

Vanderveken, D. (1990). *Meaning and Speech Acts*. Cambridge University Press.

Vanpé, A., & Aubergé, V. (2010). Prosodie expressive audio-visuelle de l'interaction personne-machine. *Technique et science informatiques, numéro spécial Agents Conversationels Animés*, *29* (in press).

Vasilescu, I., Adda-Decker, M., & Nemoto, R. (2008). Caractéristiques acoustiques et prosodiques des hésitations vocaliques dans trois langues. *TAL*, *49*(3), 199-228.

Velten, E. (1968). A laboratory task for induction of mood states. *Behaviour Research and Therapy*, *6*(4), 473-482.

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, *27*(12), 1743-1759.

Viviani, P., & Stucchi, N. (1989). The effect of movement velocity on form perception: Geometric illusions in dynamic displays. *Perception & Psychophysics,*, *46*, 266- 274.

Viviani, P., & Stucchi, N. (1992). Biological movements look uniform: evidence of motor-perceptual interactions. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 603-623.

Von Ehrenfels, C. (1890). Über Gestaltqualitäten. *Vierteljahresschrift für wissenschaftliche Philosophie*, *14*, 259–270.

Wallbott, H. G., & Riccibitti, P. (1993). Decoders Processing of Emotional Facial Expression - a Top-Down or Bottom-up Mechanism. *European Journal of Social Psychology*, *23*(4), 427-443.

Wallon, H. (1934). *Les Origines du Caractère chez l'Enfant*. Paris: Boivin.

Wallon, H. (1938). *La vie mentale*. Paris: Editions sociales.

Ward, N. (1991). Understanding spontaneous speech: the phoenix system. In *Proceedings of ICASSP* (p. 365–367). Toronto.

Ward, N. (2000). Issues in the Transcription of English Conversational Grunts. In *1st SIGdial Workshop on Discourse and Dialogue. ACL*.

Ward, N. (2000). The Challenge of Non-lexical Speech Sounds. In *ICSLP-2000* (Vol. 2, pp. 571-574). Beijing, China.

Ward, N. (2004). Pragmatic Functions of Prosodic Features in Non-Lexical Utterances. In *Speech prosody* (pp. 325-328).

Ward, N. (2006). Non-lexical conversational sounds in American English. *Pragmatics & Cognition*, *14*(1), 129-182.

Ward, N., Rivera, A., Ward, K., & Novick, D. (2005). Some Usability Issues and Research Priorities in Spoken Dialog Applications.

Wehrle, T., Kaiser, S., Schmidt, S., & Scherer, K. R. (2000). Studying the dynamics of emotional expression using synthesized facial muscle movements. *Journal of Personality and Social Psychology*, *78*(1), 105-119. US: American Psychological Association.

Weil-Barais, A. (2001). *L'homme cognitif*. Paris: Presses Universitaires de France.

Wharton, T. (2003). Interjections, language, and the 'showing/saying' continuum. *Pragmatics & Cognition*, *11*(1), 39-91. John Benjamins Publishing Company.

Wichmann, A. (2002). Attitudinal Intonation and the Inferential Process. In *Speech Prosody 2002*. Aix-en-Provence.

Wierzbicka, A. (1992). The semantics of interjection. *Journal of Pragmatics*, *18*, 159-192.

Wilkins, D. (1992). Interjections as deictics. *Journal of Pragmatics*, *18*, 119-158.

Wilting, J., Krahmer, E. J., & Swerts, M. (2006). Real vs. acted emotional speech. In *INTERSPEECH 2006 - ICSLP. Ninth International Conference on Spoken Language Processing* (p. paper 1093). Pittsburgh, PA, USA.

Woodworth, R. (1938). *Experimental Psychology*. New York: Holt.

Wranik, T. (2009). La Personnalité Des Emotions. In D. Sander & K. Scherer, *Traité de psychologie des émotions* (Dunod., pp. 359-382). Paris.

Wundt, W. (1874). *Grundzüge de physiologischen Psychologie*. Leipzig: Engelman.

Wundt, W. (1900). *Völkerpsychologie. Vol. 1. Die Sprache*. Leipzig: Engelmann.

Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society* (pp. 567-577). Chicago: Chicago Linguistic Society.

# TABLE DES ILLUSTRATIONS

## I. Figures

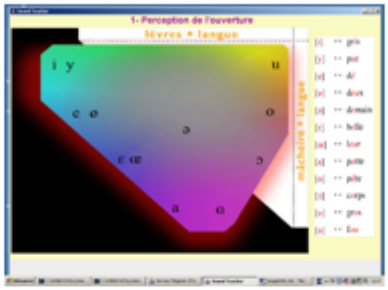
# II. Tableaux

# ANNEXES

## *Annexe 1 : Extrait du premier étiquetage linéaire du sujet F_S*

Les auto-annotations du sujet sont surlignées en jaune.

"- A 3mn48 : « page + suivante » BIP (3mn49) : IGLe descend sa lèvre inf sur la gauche en la mordant sur la droite de 3mn50 à 3mn53 et IGSour6 fronce les sourcils, lit




- A 4mn08 : « page suivante » BIP (4mn10) IGRe6 regarde en bas à droite à 4mn11

Perception ouverture: "grande concentration, le but étant de comprendre ce qui est prononcé"

*Groupe de stimuli 1*

- Sti (4mn15) : « euh » (4mn18) IGRe1 regarde vers le haut, IGSour1 lève le sourcil gauche, IGLe relève angle de la bouche pincée à 4mn20 et dit « jaune » (4mn21) IGS8 légèrement et IGY2 plisse les yeux à 4mn24

- Sti (4mn25) IGY2 plisse les yeux puis IGLe descend sa lèvre inf sur la gauche en la mordant sur la droite à 4mn28 : « vert » (4mn29)

- Sti (4mn33) IGY2 plisse les yeux en même temps : « vert » (4mn34)

- Sti (4mn38) : « vert » (4mn39)

- Sti (4mn42) : « jaune » (4mn43)

- Sti (4mn47) : entrouvre la bouche comme pour se préparer à parler « jaune » IGTe1 en hochant légèrement la tête (4mn49) puis IGLe4 étire l'extrémité droite de sa bouche

- Sti (4mn52) : « vert » (4mn53) puis IGLe5 étire légèrement et furtivement les lèvres

## *Annexe 2 : Fichier de spécification ANVIL pour les micro-expressions de la face*

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<annotation-spec>
 <head>
  <valuetype-def>
   <valueset name="tache">
      <value-el>pendant lecture</value-el>
      <value-el>pendant page suivante</value-el>
      <value-el>apres page suivante</value-el>
      <value-el>sur stimulus</value-el>
      <value-el>pendant prononciation</value-el>
      <value-el>entre prononciation</value-el>
      <value-el>pendant commentaires</value-el>
      <value-el>apres commentaires</value-el>
      <value-el>pendant reponse</value-el>
      <value-el>avant reponse</value-el>
      <value-el>apres reponse</value-el>
   </valueset>
  </valuetype-def>
  <valuetype-def>
   <valueset name="symétrie">
      <value-el>symétrique</value-el>
      <value-el>à gauche</value-el>
      <value-el>à droite</value-el>
      <value-el>surtout à gauche</value-el>
      <value-el>surtout à droite</value-el>
      <value-el>non-lieu</value-el>
   </valueset>
  </valuetype-def>
  <valuetype-def>
   <valueset name="intensité">
      <value-el>très légère</value-el>
      <value-el>légère</value-el>
      <value-el>assez légère</value-el>
      <value-el>normal</value-el>
      <value-el>assez forte</value-el>
      <value-el>forte</value-el>
      <value-el>très forte</value-el>
   </valueset>
  </valuetype-def>
  <valuetype-def>
   <valueset name="répétition">
```

```
        <value-el>1 fois</value-el>
        <value-el>plusieurs fois</value-el>
     </valueset>
   </valuetype-def>
  </head>
  <body>
   <track-spec name="audio" type="waveform" /> <!-- importation du "wave" -->
   <set-spec name="info_fichier d'annotation" >
         <attribute name="contenu" valuetype="String" />
         <attribute name="début" valuetype="String" />
         <attribute name="fin" valuetype="String" />
   </set-spec>
   <track-spec name="speech" type="primary">    <!-- transcription du verbal -->
    <attribute name="contenu" valuetype="String" />
   </track-spec>
   <track-spec name="auto-annotation" type="primary">
    <attribute name="contenu" valuetype="String" />
   </track-spec>
   <track-spec name="taches" type="primary">
    <attribute name="tache" valuetype="tache"/>
    <attribute name="tour de parole" valuetype="ReciprocalLink(TP)"/>
   </track-spec>
   <track-spec name="interjection" type="primary">
    <attribute name="type" valuetype="String" />
   </track-spec>
   <track-spec name="bruit de bouche" type="primary">
    <attribute name="type" valuetype="String" />
   </track-spec>
   <group name="Icone gestuelle" collapse="true">
    <track-spec name="IGSour" type="primary" color-attr="type">
<doc>IGSour (sourcils)<br/>IGSour1 : hausse sourcil(s)<br/>IGSour2 : fronce sourcil(s)<br/></doc>
      <attribute name="type" valuetype="Number(1,2)" >
        <value-el color="dark blue">1</value-el>
       <value-el color="blue">2</value-el>
      </attribute>
         <attribute name="intensité" valuetype="intensité" emptyvalue="false" defaultvalue="normal" />
         <attribute name="symétrie" valuetype="symétrie" emptyvalue="false" defaultvalue="symétrique" />
      <attribute name="répétition" valuetype="répétition" />
      <attribute name="furtive" valuetype="Boolean" defaultvalue="false" />
      <attribute name="stable" valuetype="Boolean" defaultvalue="true" />
    </track-spec>
   <track-spec name="IGY" type="primary" color-attr="type">
```

<doc>IGY (yeux)<br/>IGY1 : écarquille 1 oeil/les yeux<br/>IGY2 : plisse 1 oeil/les yeux<br/>IGY3 : ferme 1 oeil/les yeux<br/>IGY4 : cligne les yeux<br/>IGY5 : papillonne des yeux<br/></doc>
<attribute name="type" valuetype="Number(1,5)" >
<value-el color="light blue">1</value-el>
<value-el color="dark cyan">2</value-el>
<value-el color="cyan">3</value-el>
<value-el color="light cyan">4</value-el>
<value-el color="light cyan">5</value-el>
</attribute>
<attribute name="intensité" valuetype="intensité" emptyvalue="false" defaultvalue="normal" />
<attribute name="furtive" valuetype="Boolean" defaultvalue="false" />
<attribute name="symétrie" valuetype="symétrie" emptyvalue="false" defaultvalue="symétrique" />
</track-spec>
<track-spec name="**IGRe**" type="primary" >
<doc>IGRe (regards)<br/>IGRe1 : haut<br/>IGRe2 : haut / gauche<br/>IGRe3 : gauche<br/>IGRe4 : bas / gauche<br/>IGRe5 : bas<br/> IGRe6 : bas / droite<br/>IGRe7 : droite<br/>IGRe8 : haut / droite<br/>IGRe9 : sur la caméra<br/></doc>
<attribute name="type" valuetype="Number(1,9)" />
<attribute name="furtive" valuetype="Boolean" defaultvalue="false" />
<attribute name="netteté" valuetype="Boolean" defaultvalue="true" />
<attribute name="stable" valuetype="Boolean" defaultvalue="true" />
</track-spec>
<track-spec name="**IGN**" type="primary">
<doc>IGN (nez)<br/>IGN1 : fronce le nez<br/>IGN2 : dilate les narines<br/>IGN3 : remonte l'aile/les ailes du nez<br/></doc>
<attribute name="type" valuetype="Number(1,3)" />
<attribute name="intensité" valuetype="intensité" emptyvalue="false" defaultvalue="normal" />
<attribute name="furtive" valuetype="Boolean" defaultvalue="false" />
<attribute name="symétrie" valuetype="symétrie" emptyvalue="false" defaultvalue="symétrique" />
</track-spec>
<track-spec name="**IGTe**" type="primary">
<doc>IGTe (tête et cou)<br/>IGTe1 : hoche la tête<br/>IGTe2 : avance la tête<br/>IGTe3 : recule la tête<br/>IGTe4 : secoue la tête latéralement<br/>IGTe5 : secoue la tête (type "non")<br/>IGTe6 : penche la tête<br/>IGTe7 : mouvement instable de la tête vers l'arrière<br/>IGTe8 : grandit son cou<br/>IGTe9 : tasse son cou</doc>
<attribute name="type" valuetype="Number(1,9)" />
<attribute name="intensité" valuetype="intensité" emptyvalue="false" defaultvalue="normal" />

<attribute name="symétrie" valuetype="symétrie" emptyvalue="false" defaultvalue="non-lieu" />
<attribute name="furtive" valuetype="Boolean" defaultvalue="false" />
<attribute name="répétition" valuetype="répétition" />
<attribute name="stable" valuetype="string" defaultvalue="oui" >
<value-el>oui</value-el>
<value-el>non</value-el>
<value-el>faible</value-el>
</attribute>
</track-spec>
<track-spec name="**IGLe**" type="primary"> <!-- champ principal lèvres -->
<attribute name="intensité" valuetype="intensité" emptyvalue="false" defaultvalue="normal" />
<attribute name="symétrie" valuetype="symétrie" emptyvalue="false" defaultvalue="symétrique" />
<attribute name="furtive" valuetype="Boolean" defaultvalue="false" />
<attribute name="stable" valuetype="string" defaultvalue="oui" >
<value-el>oui</value-el>
<value-el>non</value-el>
<value-el>faible</value-el>
</attribute>
<attribute name="ouverture bouche" >
<value-el>fermée, lèvres pincées</value-el>
<value-el>fermée</value-el>
<value-el>entrouverte</value-el>
<value-el>ouverte</value-el>
</attribute>
</track-spec>
<doc> **Sous-catégories de paramètres au niveau des lèvres**</doc>
<track-spec name="**IGLe-**" type="singleton" ref="Icone gestuelle.IGLe" >
<doc>IGLe-<br/>IGLe-1 : bouche en cul de poule<br/>IGLe-2 : lèvres rentrées<br/>IGLe-3 : étire sa lèvre inférieure d'un côté<br/>IGLe-4 : descend sa lèvre inférieure<br/></doc>
<attribute name="type" valuetype="Number(1,4)" />
</track-spec>
<track-spec name="**IGLeAng**" type="singleton" ref="Icone gestuelle.IGLe" >
<doc>IGLeAng<br/>IGLeAng1 : angles des lèvres relevés<br/>IGLeAng2 : angles des lèvres abaissés<br/></doc>
<attribute name="type" valuetype="Number(1,2)" />
</track-spec>
<track-spec name="**IGLeArr**" type="singleton" ref="Icone gestuelle.IGLe" >
<doc>IGLeArr<br/>IGLeArr1 : lèvres protruses<br/>IGLeArr2 : lèvres étirées<br/></doc>
<attribute name="type" valuetype="Number(1,2)" />
</track-spec>

```
      <track-spec name="IGLeD" type="singleton" ref="Icone gestuelle.IGLe" >
<doc>IGLeD (dents)<br/>IGLeD1 : dents visibles<br/>IGLeD2 : dents inf visibles<br/></doc>
        <attribute name="type" valuetype="Number(1,2)" />
      </track-spec>
    <track-spec name="IGJ1" type="primary">
      <doc>IGJ (joues)<br/>IGJ1 : joues gonflées<br/></doc>
          <attribute name="intensité" valuetype="intensité" emptyvalue="false" defaultvalue="normal" />
      <attribute name="furtive" valuetype="Boolean" defaultvalue="false" />
     </track-spec>
    <track-spec name="IGMe" type="primary">
<doc>IGMe (menton)<br/>IGMe1 : menton froncé<br/>IGMe2 : mâchoire abaissée<br/> </doc>
      <attribute name="type" valuetype="Number(1,2)" />
          <attribute name="intensité" valuetype="intensité" emptyvalue="false" defaultvalue="normal" />
      <attribute name="furtive" valuetype="Boolean" defaultvalue="false" />
     </track-spec>
        <group name="Idiosynchrasique" collapse="true">
                <attribute name="contenu" valuetype="String" emptyvalue="false" defaultvalue="Sabrina" />
                <track-spec name="IGMe3" type="primary" >
                        <doc>IGMe3 : bosse sur le menton<br/></doc>
                                <attribute name="intensité" valuetype="intensité" emptyvalue="false" defaultvalue="normal" />
                </track-spec>
                <track-spec name="IGFo" type="primary">
                        <doc>IGFo : fossette (surtout à droite)<br/></doc>
                                <attribute name="intensité" valuetype="intensité" emptyvalue="false" defaultvalue="normal" />
                </track-spec>
        </group>
   </group>
<group name="Physio" collapse="true"> <!-- signaux physiologiques. Exemple de champ : -->
 <group name="Synchro">
<track-spec name="synchro" type="multiplot" height="1" has-zero-axis="true"/>
</group> <!-- etc. -->
</group>
</body>
```

## *Annexe 3 : Correspondance partielle entre les FACS et nos icônes*

(issus de Loyau, 2007, p.193)

| N° AU | FACS | Traduction | Muscles recrutés | Equivalence avec notre codage |
|---|---|---|---|---|
| 1 | Inner Brow Raiser | "élevation sourcil interne" | Frontalis, Pars Medialis | |
| 2 | Outer Brow Raiser | "élevation sourcil externe" | Frontalis, Pars Lateralis | IGSour1 |
| 4 | Brow Lowerer | "abaissement du sourcil" | Depressor Glabellae ; Depressor Supercilli ; Corrugator | IGSour2 |
| 5 | Upper Lid Raiser | "élevation de la paupière supérieure" | Levator Palpebrae Superioris | IGY1 |
| 6 | Cheek Raiser | "élevation de la joue" | Orbicularis Oculi, Pars Orbitalis | |
| 7 | Lid Tightener | "plissement, resserrement des paupières" | Orbicularis Oculi, Pars Palebralis | |
| 8 | Lips Toward Each Other | "pincement / fermeture des lèvres" | Orbicularis Oris | |
| 9 | Nose Wrinkler | "plissement du nez" | Levator Labii Superioris, Alaeque Nasi | IGN1 |
| 10 | Upper Lip Raiser | "élevation de la lèvre supérieure" | Levator Labii Superioris, Caput Infraorbitalis | |
| 11 | Nasolabial Furrow Deepener | "creusement du sillon nasolabial" | Zygomatic Minor | IGLe-1 |
| 12 | Lip Corner Puller | "tirer/presser le coin des lèvres" | Zygomatic Major | |
| 13 | Cheek Puffer | "gonflement des joues" | Caninus | |
| 14 | Dimpler | "fossettes" | Buccinnator | |
| 15 | Lip Corner Depressor | "abaissement du coin des lèvres" | Triangularis | IGLeAng2 |
| 16 | Lower Lip Depressor | "abaissement de la lèvre inférieure" | Depressor Labii | IGLe-4 |
| 17 | Chin Raiser | "élevation du menton" | Mentalis | |
| 18 | Lip Puckerer | "avancement des lèvres" | Incisivii Labii Superioris; Incisivii Labii Inferioris | IGLeArr1 |

| | | | | |
|---|---|---|---|---|
| 20 | Lip Stretcher | "étirement des lèvres" | Risorius | IGLeArr2 |
| 22 | Lip Funneler | "protrusion / lèvres en entonnoir" | Orbicularis Oris | IGLe-1 |
| 23 | Lip Tightner | "plissement, resserrement des lèvres" | Orbicularis Oris | variable "lèvres pincées" |
| 24 | Lip Pressor | "pincement des lèvres" | Orbicularis Oris | |
| 25 | Lips Part | "ouverture / écartement des lèvres" | Depressor Labii, or Relaxation of Mentalis or Orbicularis Oris | |
| 26 | Jaw Drop | "abaissement de la mâchoire" | Masetter ; Temporal and Internal Pterygoid Relaxed | IGMe2 |
| 27 | Mouth Stretch | "étirement de la bouche" | Pterygoids; Digastric | |
| 28 | Lip Suck | "sucer / aspirer / têter" | Orbicularis Oris | |
| 38 | Nostril Dilator | "dilatation des narines" | Nasalis, Pars Alaris | IGN2 |
| 39 | Nostril Compressor | "compression des narines" | Nasalis, Pars Transversa and Depressor Septi Nasi | |
| 41 | Lid Droop | "abaissement des paupières" | Relaxation of Levator Palpebrae Superioris | |
| 42 | Slit | "loucher / jeter un coup d'œil" | Orbicularis Oculi | IGRe + numéro |
| 43 | Eyes Closed | "fermeture des yeux" | Relaxation of Levator Palpebrae Superioris | IGY3 |
| 44 | Squint | "plissement des yeux" | Orbicularis Oculi, Pars Palpebralis | IGY2 |
| 45 | Blink | "clignement des yeux (paupières mi-fermées)" | Relaxation of Levator Palpebrae and Contraction of Orbicularis Oculi, Pars Palpebralis | IGY5 |
| 46 | Wink | "clignement des yeux" | Orbicularis Oculi | IGY4 |

## *Annexe 4 : Différences entre auto-annotations données par les sujets et labels proposés aux juges, accompagnées pour chaque auto-annotation de sa place approximative dans le scénario*

### Sujet S

| Auto-annotation | Phase | Label utilisé | Temps de début |
|---|---|---|---|
| "'emprise' du logiciel dans le sens où je suis les consignes du mieux que je peux" | 1 | "emprise" du logiciel dans le sens où elle suit les consignes du mieux qu'elle peut | 8,03 |
| "'emprise' du logiciel dans le sens où je suis les consignes du mieux que je peux" | 2 | "emprise" du logiciel dans le sens où elle suit les consignes du mieux qu'elle peut | 11,04 |
| "au pif, une envie de rigoler" | 4 | envie de rigoler et répond au hasard | 32,49 |
| "au pif, une envie de rigoler" | 4 | envie de rigoler et répond au hasard | 32,55 |
| "concentration mais réponses parfois hasardeuses, la tâche devient répétitive et ennuyeuse" | 3 | concentrée et répond au hasard | 24,33 |
| "réponses avec une pointe de sérieux et beaucoup d'approximations" | 4 | concentrée et répond au hasard | 35,19 |
| "grande concentration, le but étant de comprendre ce qui est prononcé" | 1 | concentrée | 4,28 |
| "retour concentration" | 4 | concentrée | 30,49 |
| "déçue par les résultats, j'essaie de trouver des solutions" | 4 | déçue | 28,4 |
| "déçue par les résultats, j'essaie de trouver des solutions" | 4 | déçue | 28,55 |
| "écoute attentive" | 2 | écoute attentivement | 14,38 |
| "écoute attentive" | 2 | écoute attentivement | 14,43 |
| "un petit peu de mal à commencer, une certaine envie de rigoler" | 1 | pas concentrée et envie de rigoler | 2,58 |
| "un petit peu de mal à commencer, une certaine envie de rigoler" | 1 | pas concentrée et envie de rigoler | 3,08 |
| "je commence un peu à « rire jaune » de mes résultats" | 4 | "rit jaune" de ses résultats | 31,31 |
| "je commence un peu à « rire jaune » de mes résultats" | 4 | "rit jaune" de ses résultats | 31,56 |
| léger stress qui apparaît | 1 | stressée | 5,48 |
| "la difficulté s'accroît donc le stress aussi" | 1 | stressée | 6,43 |

## Sujet T

| Auto-annotation | Phase | Label utilisé | début |
|---|---|---|---|
| "calme ?" | 1 | calme / va bien | 10,54 |
| "calme" | 3 | calme / va bien | 22,16 |
| "ça va..." | 4 | calme / va bien | 39,11 |
| "ça va..." | 4 | calme / va bien | 40,12 |
| "concentration" | 1 | concentrée | 6,09 |
| "concentrée" | 2 | concentrée | 11,06 |
| "concentrée" | 2 | concentrée | 11,13 |
| "concentrée" | 2 | concentrée | 11,34 |
| "j'ai l'air déçue" | 1 | déçue | 7,31 |
| "déception" | 3 | déçue | 23,38 |
| "déception" | 3 | déçue | 30,17 |
| "déception" | 3 | déçue | 30,37 |
| "étonnée ?" | 1 | étonnée | 8,44 |
| "étonnée ?" | 1 | étonnée | 9,12 |
| "étonnée ?" | 1 | étonnée | 9,17 |
| "étonnée" | 4 | étonnée | 38,02 |
| "hésitation" | 3 | hésitante | 22,03 |
| "hésitation" | 3 | hésitante | 22,21 |
| "hésitante" | 4 | hésitante | 36,3 |
| "hésitante" | 4 | hésitante | 37,08 |
| "pas à l'aise, inquiète" | 1 | mal à l'aise / inquiète | 3,12 |
| "mal à l'aise" | 2 | mal à l'aise / inquiète | 13,06 |
| "mal assurée" | 3 | mal à l'aise / inquiète | 26,06 |
| "mal assurée" | 3 | mal à l'aise / inquiète | 26,15 |
| "oppressée (suis une grande stressée)" | 2 | angoissée / oppressée | 13,56 |
| "oppressée (suis une grande stressée)" | 2 | angoissée / oppressée | 13,57 |
| "angoissée ?" | 3 | angoissée / oppressée | 28,44 |
| "angoissée ?" | 3 | angoissée / oppressée | 28,58 |
| "rassurée, plus détendue" | 2 | rassurée / plus détendue | 19,12 |
| "rassurée, plus détendue" | 2 | rassurée / plus détendue | 20,05 |
| "rassurée, plus détendue" | 2 | rassurée / plus détendue | 20,08 |
| "rassurée, plus détendue" | 2 | rassurée / plus détendue | 20,09 |
| "stress" | 2 | stressée | 16,32 |
| "stress" | 2 | stressée | 17,0788 |
| "stress" | 2 | stressée | 18 |
| "stress (encore!!)" | 3 | stressée | 29,28 |
| "perplexité" | 1 | un peu perdue / perplexe | 5,31 |
| "un peu perdue" | 1 | un peu perdue / perplexe | 6,53 |
| "un peu perdue" | 1 | un peu perdue / perplexe | 6,59 |
| "un peu perdue" | 1 | un peu perdue / perplexe | 7,33 |

# *Annexe 5 : Liste des stimuli du test et de leurs caractéristiques*

| **Sujet S** | **Stimuli dynamiques** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Label | Date de début en seconde | Date de fin en seconde | Durée en seconde | Instant en min. de la vidéo pour le stim. Statique | Auto-annotation | Description au niveau de la bouche | Haut du visage bouge ? | Mouvement tête, cou et/ou buste? |
| "« emprise » du logiciel" | Ph.1: 482,2 | 483,44 | 1,24 | 8,03 | "« emprise » du logiciel dans le sens où je suis les consignes du mieux que je peux" | bouche en cul de poule / protrusion | | |
| "« emprise » du logiciel" | Ph.2: 53,88 | 55,04 | 1,16 | 11,04 | "« emprise » du logiciel dans le sens où je suis les consignes du mieux que je peux" | descend lèvre inf sur la gauche en la mordant sur la droite | oui | |
| "au pif, une envie de rigoler" | Ph.4: 328,88 | 329,92 | 1,04 | 32,49 | "au pif, une envie de rigoler" | relève angles de la bouche entrouverte | | oui |
| "au pif, une envie de rigoler" | Ph.4: 334,44 | 336,08 | 1,64 | 32,55 | "au pif, une envie de rigoler" | relève angles de la bouche ouverte | | |
| "concentration + hasard" | Ph.3: 372,36 | 373,16 | 0,8 | 24,33 | "concentration mais réponses parfois hasardeuses, la tâche devient répétitive et ennuyeuse" | relève angles des lèvres pincées | | |
| "concentration + hasard" | Ph.4: 479,32 | 480,12 | 0,8 | 35,19 | "réponses avec une pointe de sérieux et beaucoup d'approximations" | bouche en cul de poule / protrusion | oui | |
| "concentration" | Ph.1: 268,2 | 269,28 | 1,08 | 4,28 | "grande concentration, le but étant de comprendre ce qui est prononcé" | descend lèvre inf sur la gauche | oui | |
| "concentration" | Ph.4: 209,24 | 209,56 | 0,32 | 30,49 | "retour concentration" | descend lèvre inf sur la gauche | | oui |
| "déçue" | Ph.4: 80,24 | 80,64 | 0,4 | 28,4 | "déçue par les résultats, j'essaie de trouver des solutions" | relève angles des lèvres pincées | oui | |
| "déçue" | Ph.4: 95,48 | 96,92 | 1,44 | 28,55 | "déçue par les résultats, j'essaie de trouver des solutions" | bouche en cul de poule / protrusion | oui | |
| "écoute attentive" | Ph.2: 268,76 | 268,8 | 0,04 | 14,38 | "écoute attentive" | étire angles des lèvres | | |
| "écoute attentive" | Ph.2: 273,64 | 273,68 | 0,04 | 14,43 | "écoute attentive" | étire angles des lèvres | | |
| "pas concentrée, envie de rigoler" | Ph.1: 178,28 | 179,2 | 0,92 | 2,58 | "un petit peu de mal à commencer, une certaine envie de rigoler" | descend lèvre inf sur la gauche | oui | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| "pas concentrée, envie de rigoler" | Ph.1: 187,68 | 188,32 | 0,64 | 3,08 | "un petit peu de mal à commencer, une certaine envie de rigoler" | relève angles de la bouche fermée ou entrouverte | | |
| "rit jaune de ses résultats" | Ph.4: 251,6 | 252,68 | 1,08 | 31,31 | "je commence un peu à « rire jaune » de mes résultats" | étire lèvre inf sur la gauche | oui | |
| "rit jaune de ses résultats" | Ph.4: 276,44 | 277,2 | 0,76 | 31,56 | "je commence un peu à « rire jaune » de mes résultats" | relève angles de la bouche ouverte | | |
| "stress" | Ph.1: 348,16 | 349,08 | 0,92 | 5,48 | léger stress qui apparaît | étire lèvre inf sur la gauche | oui | oui |
| "stress" | Ph1: 402,4 | 403,44 | 1,04 | 6,43 | "la difficulté s'accroît donc le stress aussi" | bouche en cul de poule / protrusion | | |
| | | | | | | | | |
| **Sujet T** | **Stimuli dynamiques** | | | | | | | |
| "mal à l'aise"/"inquiète" | 192,2 | 193,64 | 1,44 | 3,12 | "pas à l'aise, inquiète" | relève angle droit de la bouche fermée, lèvres pincées et étirées | oui | |
| "un peu perdue" / "perplexe" | 330,92 | 331,48 | 0,56 | 5,31 | "perplexité" | | oui | oui |
| "concentrée" | 370 | 371,48 | 1,48 | 6,09 | "concentrée" | relève angles de la bouche fermée, lèvres pincées | | oui |
| "un peu perdue" / "perplexe" | 412,56 | 413,68 | 1,12 | 6,53 | "un peu perdue" | descend angles de la bouche fermée | oui | |
| "un peu perdue" / "perplexe" | 419,48 | 419,92 | 0,44 | 6,59 | "un peu perdue" | | oui | oui |
| "déçue" | 451,04 | 451,72 | 0,68 | 7,31 | "air déçu" | relève angles de la bouche entrouverte, mais s'apprête à parler | oui | oui |
| "un peu perdue" / "perplexe" | 453,28 | 453,68 | 0,4 | 7,33 | "un peu perdue" | lèvres étirées | oui | oui |
| "étonnée" | 524,28 | 526,72 | 2,44 | 8,44 | "étonnée ?" | relève angle droit de la bouche fermée, lèvres étirées | oui | oui |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| "étonnée" | 552,72 | 553,04 | 0,32 | 9,12 | "étonnée ?" | relève angle droit de la bouche ouverte | oui | oui |
| "étonnée" | 556,44 | 558,6 | 2,16 | 9,17 | "étonnée ?" | relève angles de la bouche fermée, lèvres pincées et étirées | oui | oui |
| "calme"/"ça va" | 654,84 | 655,2 | 0,36 | 10,54 | "calme" | bouche fermée, légèrement protruse | | oui |
| "concentrée" | 665,48 | 665,96 | 0,48 | 11,06 | "concentrée" | | | oui |
| "concentrée" | 671,84 | 673,32 | 1,48 | 11,13 | "concentrée" | | oui | oui |
| "concentrée" | 694,64 | 695,28 | 0,64 | 11,34 | "concentrée" | | oui | oui |
| "mal à l'aise"/"inquiète" | 785,96 | 786,88 | 0,92 | 13,06 | "mal à l'aise" | bouche protruse, lèvres pincées et étirées | oui | oui |
| "oppressée, angoissée" | 836,24 | 838,36 | 2,12 | 13,56 | "oppressée" | relève angles de la bouche légèrement entrouverte, lèvres étirées | oui | oui |
| "oppressée, angoissée" | 836,24 | 838,36 | 2,12 | 13,57 | "oppressée" | relève angles de la bouche, lèvres étirées et pincées | oui | oui |
| "stressée" | 992,6 | 993,44 | 0,84 | 16,32 | "stressée" | rentre les lèvres étirées, bouche entrouverte | oui | oui |
| "stressée" | 1027,84 | 1028 | 0,16 | 17,0788 | "stressée" | bouche entrouverte | oui | oui |
| "stressée" | 1079,72 | 1081,4 | 1,68 | 18 | "stressée" | bouche légèrement protruse, entrouverte | oui | oui |
| "rassurée, plus détendue" | 1150,92 | 1152,84 | 1,92 | 19,12 | "rassurée, plus détendue" | étire angles des lèvres pincés | oui | oui |
| "rassurée, plus détendue" | 1205,72 | 1205,8 | 0,08 | 20,05 | "rassurée, plus détendue" | relève angles de la bouche fermée, lèvres étirées | oui | oui |
| "rassurée, plus détendue" | 1207,92 | 1208,28 | 0,36 | 20,08 | "rassurée, plus détendue" | relève angles de la bouche fermée | | |
| "rassurée, plus détendue" | 1208,68 | 1210,16 | 1,48 | 20,09 | "rassurée, plus détendue" | relève angles de la bouche fermée, lèvres pincées et étirées | | |
| "hésitante" | 1323,68 | 1325,04 | 1,36 | 22,03 | "hésitante" | relève angles de la bouche ouverte, lèvres étirées, surtout à droite | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| "calme"/"ça va" | 1334,24 | 1338,08 | 3,84 | 22,16 | "calme" | relève angles de la bouche fermée | | oui |
| "hésitante" | 1341,52 | 1343,08 | 1,56 | 22,21 | "hésitante" | bouche protruse et lèvres pincées vers la droite | oui | oui |
| "déçue" | 1418,08 | 1419 | 0,92 | 23,38 | "déception" | relève angles de la bouche fermée | oui | |
| "mal à l'aise"/"inquiète" | 1565,52 | 1569,28 | 3,76 | 26,06 | "mal assurée" | relève angle droit de la bouche légèrement entrouverte | oui | oui |
| "mal à l'aise"/"inquiète" | 1573,84 | 1576,36 | 2,52 | 26,15 | "mal assurée" | descend angles de la bouche fermée | oui | oui |
| "oppressée, angoissée" | 1723,88 | 1724,56 | 0,68 | 28,44 | "angoissée" | relève angles de la bouche, lèvres étirées | oui | |
| "oppressée, angoissée" | 1738,6 | 1738,88 | 0,28 | 28,58 | "angoissée" | relève angles de la bouche légèrement entrouverte, lèvres étirées | oui | oui |
| "stressée" | 1768,92 | 1770,8 | 1,88 | 29,28 | "stressée" | relève angles de la bouche, lèvres rentrées, pincées et étirées | oui | oui |
| "déçue" | 1816,92 | 1818,04 | 1,12 | 30,17 | "déception" | relève angles de la bouche fermée, lèvres pincées et étirées | oui | |
| "déçue" | 1837,48 | 1838,24 | 0,76 | 30,37 | "déception" | relève angles de la bouche fermée, lèvres étirées | oui | |
| "hésitante" | 2189,56 | 2191,24 | 1,68 | 36,3 | "hésitante" | bouche protruse et lèvres pincées | oui | oui |
| "hésitante" | 2228,32 | 2229,28 | 0,96 | 37,08 | "hésitante" | relève angles de la bouche légèrement entrouverte | oui | |
| "étonnée" | 2280,2 | 2282,92 | 2,72 | 38,02 | Warning final | relève angles de la bouche fermée, lèvres pincées | | oui |
| "calme"/"ça va" | 2350,56 | 2352,08 | 1,52 | 39,11 | "ça va" | relève angles de la bouche fermée | | oui |
| "calme"/"ça va" | 2411,92 | 2414,24 | 2,32 | 40,12 | "ça va" | relève angle de la bouche fermée, lèvres étirées | | |

## *Annexe 6 : Scripts modifiés, qui définissent l'interface et permettent de présenter les stimuli, tout en enregistrant automatiquement les réponses des juges sous forme de fichiers texte*

Précision : en regard de la différence de traitement de différents caractères entre PC et Macintosh, ainsi que sous Matlab, mais également en vu des futurs traitements statistiques sur les résultats du test, il est important d'uniformiser et d'éviter caractères spéciaux (lettres accentuées…) et espace, à la fois pour les noms attribués aux différents objets de l'interface et pour les noms des fichiers de stimuli.

**Script du bouton *init* permettant d'initialiser le test pour chacun des sujets** :

```
-- on designe comme dossier par defaut le dossier contenant la liste des stimuli
set the itemDel to "/"
put the number of items in fichStimuli into nb
put 1 into i
put "" into repCourant
repeat for nb-1 times
  put item i of fichStimuli after repCourant
  put "/" after repCourant
  put i+1 into i
end repeat
set the defaultFolder to repCourant

 -- on demande les infos sur le sujet
ask "Donnez vos initiales, genre et age."
put it into nomSujet
put repCourant & "Fich_" & nomSujet & "_Tiph.txt" into fichRep
-- on ouvre le fichier de resultats
open file fichRep
write "Test Tiphaine, resultats de " & nomSujet & return to file fichRep
 write "Ordre de passage" & tab & "Fichier stimulus" & tab & "Condition" & tab & "Reponse" & return to file fichRep

end mouseUp
```

**Script du bouton *debut*, permettant de lancer le premier stimulus :**

```
on mouseUp
    global repCourant, listeStimuli, nomSujet, fichRep, stimuliCourant, numStimuliCourant, nbStimuli, condition, premiereCondition, Etiquette
```

```
-- quelques initialisations
set the itemDel to return
put the number of items of listeStimuli into nbStimuli
put 1 into numStimuliCourant
put empty into stimuliCourant

-- on deselectionne tous les boutons
unhilite button id 1899
unhilite button id 1900
unhilite button id 1901
unhilite button id 1902
unhilite button id 1903
unhilite button id 1904
unhilite button id 1905
unhilite button id 1906
unhilite button id 1907
unhilite button id 1908

-- on met a jour l'affichage
set the visible of button id 1899 to false
set the visible of button id 1900 to false
set the visible of button id 1901 to false
set the visible of button id 1902 to false
set the visible of button id 1903 to false
set the visible of button id 1904 to false
set the visible of button id 1905 to false
set the visible of button id 1906 to false
set the visible of button id 1907 to false
set the visible of button id 1908 to false

set the visible of field id 1898 to false
set the visible of graphic id 1008 to false
set the filename of image id 1913 to ""
set the visible of image id 1913 to false
set the visible of button id 1031 to false
set the visible of field id 1912 to false


set the itemDelimiter to return
get item numStimuliCourant of listeStimuli
put it into stimuliCourant

put numStimuliCourant+1 into numStimuliCourant

-- on affiche les consignes specifiques de la premiere condition
```

```
  go to card id 1002
  put stimuliCourant into condition
  if (stimuliCourant = "condition haut")
  then put "Vous allez voir maintenant uniquement le haut du visage de la personne. Cliquez sur le bouton 'Suivant' pour continuer." into field id 1912
  if (stimuliCourant = "condition bas")
  then put "Vous allez voir maintenant uniquement le bas du visage de la personne. Cliquez sur le bouton 'Suivant' pour continuer." into field id 1912
  if (stimuliCourant = "condition entier")
  then put "Vous allez voir maintenant le visage complet de la personne. Cliquez sur le bouton 'Suivant' pour continuer." into field id 1912

  -- Afficher champ changement de condition + bouton suivant
  set the visible of field id 1912 to true
  set the visible of button id 1031 to true


end mouseUp
```

**Script du bouton *suivant*, permettant d'enregistrer la réponse donnée et de passer au stimulus suivant :**

```
on mouseUp
    global repCourant, listeStimuli, nomSujet, fichRep, stimuliCourant, numStimuliCourant, nbStimuli, condition, premiereCondition, Etiquette, tempsSuivant


  --on enregistre le temps
  put the  long time into tempsSuivant

  -- on enregistre les resultats
  put empty into reponse
   if (stimuliCourant = "condition haut") or (stimuliCourant = "condition bas") or (stimuliCourant = "condition entier")
  then
    put stimuliCourant into reponse
    write condition & tab & tempsSuivant & return to file fichRep
  else
    if the hilite of button id 1899 is true
    then put the label of button id 1899 into reponse
    if the hilite of button id 1900 is true
    then put the label of button id 1900 into reponse
    if the hilite of button id 1901 is true
```

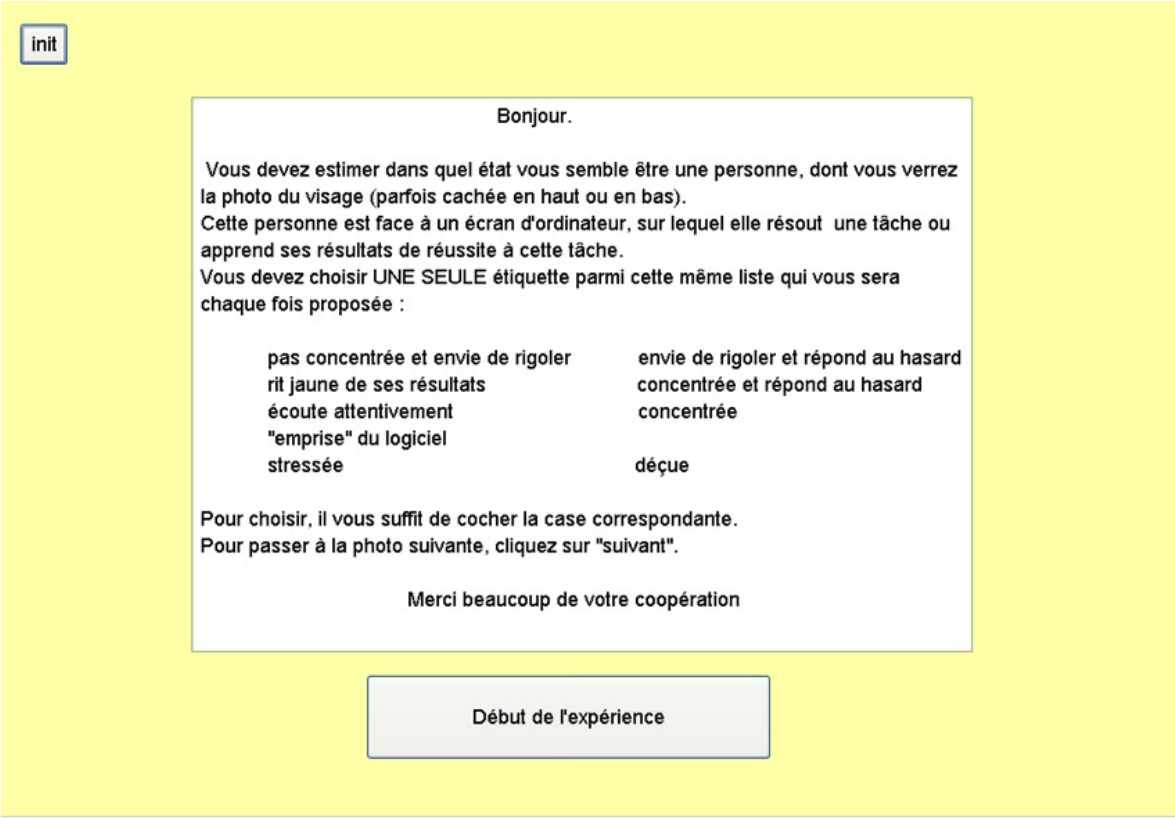
```
  then put the label of button id 1901 into reponse
  if the hilite of button id 1902 is true
  then put the label of button id 1902 into reponse
  if the hilite of button id 1903 is true
  then put the label of button id 1903 into reponse
  if the hilite of button id 1904 is true
  then put the label of button id 1904 into reponse
  if the hilite of button id 1905 is true
  then put the label of button id 1905 into reponse
  if the hilite of button id 1906 is true
  then put the label of button id 1906 into reponse
  if the hilite of button id 1907 is true
  then put the label of button id 1907 into reponse
  if the hilite of button id 1908 is true
  then put the label of button id 1908 into reponse
   write numStimuliCourant & tab & stimuliCourant & tab & condition & tab &
reponse & tab & tempsSuivant & return to file fichRep
 end if

 -- on deselectionne tous les boutons
 unhilite button id 1899
 unhilite button id 1900
 unhilite button id 1901
 unhilite button id 1902
 unhilite button id 1903
 unhilite button id 1904
 unhilite button id 1905
 unhilite button id 1906
 unhilite button id 1907
 unhilite button id 1908

 -- on met a jour l affichage
 set the visible of button id 1899 to false
 set the visible of button id 1900 to false
 set the visible of button id 1901 to false
 set the visible of button id 1902 to false
 set the visible of button id 1903 to false
 set the visible of button id 1904 to false
 set the visible of button id 1905 to false
 set the visible of button id 1906 to false
 set the visible of button id 1907 to false
 set the visible of button id 1908 to false
 set the visible of field id 1898 to false
 set the visible of graphic id 1008 to false
 set the visible of image id 1913 to false
```

```
set the visible of button id 1031 to false
set the visible of field id 1912 to false

-- si il reste des stimuli, on continue
if numStimuliCourant <= nbStimuli
then
  -- on lit le nom du stimulus suivant
  set the itemDelimiter to return
  get item numStimuliCourant of listeStimuli
  put it into stimuliCourant
  put numStimuliCourant+1 into numStimuliCourant


  -- Si on change de condition, on affiche un champ particulier
  -- Sinon, on affiche l'image suivante
   if (stimuliCourant = "condition haut") or (stimuliCourant = "condition bas") or (stimuliCourant = "condition entier")
  then
    put stimuliCourant into condition
    if (stimuliCourant = "condition haut")
     then put "Vous allez voir maintenant uniquement le haut du visage de la personne. Cliquez sur le bouton 'Suivant' pour continuer." into field id 1912
    if (stimuliCourant = "condition bas")
    then put "Vous allez voir maintenant uniquement le bas du visage de la personne. Cliquez sur le bouton 'Suivant' pour continuer." into field id 1912
    if (stimuliCourant = "condition entier")
    then put "Vous allez voir maintenant le visage complet de la personne. Cliquez sur le bouton 'Suivant' pour continuer." into field id 1912

    -- Afficher champ changement de condition + bouton suivant
    set the visible of field id 1912 to true
    set the visible of button id 1031 to true
  else
    set the visible of graphic id 1008 to true
    -- afficher l'image
    set the filename of image id 1913 to stimuliCourant
    set the visible of image id 1913 to true
    wait for 1 sec
    -- on affiche les boutons, une fois que le son est terminé
    set the visible of button id 1899 to true
    set the visible of button id 1900 to true
    set the visible of button id 1901 to true
    set the visible of button id 1902 to true
    set the visible of button id 1903 to true
    set the visible of button id 1904 to true
```

```
    set the visible of button id 1905 to true
    set the visible of button id 1906 to true
    set the visible of button id 1907 to true
    set the visible of button id 1908 to true
    set the visible of field id 1898 to true
   end if
  else
   -- on affiche un message de fin
   put "Voila, c'est fini!! Merci encore de votre participation" into field id 1912
   set the visible of field id 1912 to true
   close file fichRep
  end if
end mouseUp
```

## *Annexe 7 : Interface du test où est présentée la situation*

**Écran de présentation du test de S. :**



## *Annexe 8 : Fichier de spécification ANVIL pour l'étiquetage des événements vocaux*

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<annotation-spec>
 <head>
  <valuetype-def>
   <valueset name="tache">
      <value-el>pendant lecture</value-el>
      <value-el>pendant page suivante</value-el>
      <value-el>apres page suivante</value-el>
      <value-el>sur stimulus</value-el>
      <value-el>pendant prononciation</value-el>
      <value-el>entre prononciation</value-el>
      <value-el>pendant commentaires</value-el>
      <value-el>apres commentaires</value-el>
      <value-el>pendant reponse</value-el>
      <value-el>avant reponse</value-el>
      <value-el>apres reponse</value-el>
   </valueset>
  </valuetype-def>

  <valuetype-def>
   <valueset name="TP">
      <value-el>Avt</value-el>
      <value-el>Ap</value-el>
      <value-el>~</value-el>
      <value-el>0</value-el>
      <value-el>1</value-el>
   </valueset>
  </valuetype-def>

 </head>

 <body>

       <track-spec name="audio" type="waveform" />

       <track-spec name="speech" type="primary">
            <attribute name="contenu" valuetype="String" />
       </track-spec>

  <set-spec name="info_fichier d'annotation" >
       <attribute name="contenu" valuetype="String" />
```

```
        <attribute name="debut" valuetype="String" />
        <attribute name="fin" valuetype="String" />
  </set-spec>

<group name="PRAAT" collapse="true">
        <track-spec name="Praat Pitch" type="speech analysis"/>
        <track-spec name="Praat Intensity" type="speech analysis"/>
</group>

  <track-spec name="SousPhase" type="primary">
   <attribute name="numero" valuetype="String" />
   <attribute name="contenu" valuetype="String" />
   <attribute name="induction" valuetype="string">
     <value-el>etats mentaux</value-el>
     <value-el>sous-res</value-el>
      <value-el>resultats warning</value-el>
      <value-el>fin</value-el>
    </attribute>
  </track-spec>

  <track-spec name="phase" type="span" ref="SousPhase">
   <attribute name="phase" valuetype="Number(1,4)" />
</track-spec>

  <track-spec name="auto-annotation" type="primary">
   <attribute name="contenu" valuetype="String" />
  </track-spec>

  <track-spec name="taches" type="primary">
   <attribute name="tache" valuetype="tache"/>
   <attribute name="tour de parole" valuetype="ReciprocalLink(TP)"/>
  </track-spec>

  <track-spec name="TourDeParole" type="span" ref="taches" color-attr="TP">
   <attribute name="TP">
    <value-el color="red">pendant TP</value-el>
    <value-el color="orange">hors TP</value-el>
   </attribute>
  </track-spec>

  <track-spec name="interjection" type="primary">

     <doc>
      <b>Interj : Valeur pour Type et Type classement</b><br/>
       euh ; euh: ; [a]  = V <br/>
```

```
        euh +soupiré ; euh +chuchoté = V variante <br/>
        mmh: ; fff = C <br/>
        ouh là = VCV <br/>
        euh fff ; euh mmh = VC <br/>
        beu ; mmh heu = CV <br/>
        ben mmh ; ben: mmh ; et ben euh = Comb <br/>
    </doc>

        <attribute name="type" valuetype="String" />
        <attribute name="type classement" valuetype="String" />
        <attribute name="qualite voix" defaultvalue="modal">
                <value-el>modal</value-el>
                <value-el>soupire</value-el>
                <value-el>chuchote</value-el>
                <value-el>murmure</value-el>
                <value-el>creaky</value-el>
                <value-el>nasalise</value-el>
        </attribute>
        <attribute name="TP" valuetype="TP" defaultvalue="0"/>
</track-spec>

  <track-spec name="bruit de bouche" type="primary">

        <doc>
                <b>BB : Valeur pour Type et Type classement</b><br/>
                occlusion = relachement occlusion<br/>
                click = click ou click mouille ou click x 2<br/>
                friction<br/>
                grande inspiration (tremblante) = insp<br/>
                grande expiration = exp<br/>
                (grande) expiration brutale (voisee) = exp brutale<br/>
                avale sa salive bruyamment = avale salive<br/>
                relache sa respiration = blocage air_exp<br/>
                raclement de gorge = raclement gorge<br/>
                gemissement = gemissement<br/>
                Autres Types classement : click_insp<br/>
    </doc>

        <attribute name="classement par type element 1">
                <value-el>relachement articulateur</value-el>
                <value-el>click</value-el>
                <value-el>occlusion ingressive_insp</value-el>
                <value-el>occlusion</value-el>
                <value-el>friction</value-el>
                <value-el>insp</value-el>
```

```
        <value-el>exp</value-el>
        <value-el>exp brutale</value-el>
        <value-el>relache sa respiration</value-el>
        <value-el>langue-levres mouillees</value-el>
        <value-el>avale salive</value-el>
        <value-el>raclement de gorge</value-el>
        <value-el>gemissement</value-el>
        <value-el>autre</value-el>
</attribute>
<attribute name="classement par type element 2" defaultvalue="non lieu">
        <value-el>non lieu</value-el>
        <value-el>relachement articulateur</value-el>
        <value-el>click</value-el>
        <value-el>occlusion ingressive_insp</value-el>
        <value-el>occlusion</value-el>
        <value-el>friction</value-el>
        <value-el>insp</value-el>
        <value-el>exp</value-el>
        <value-el>exp brutale</value-el>
        <value-el>relache sa respiration</value-el>
        <value-el>langue-levres mouillees</value-el>
        <value-el>avale salive</value-el>
        <value-el>raclement de gorge</value-el>
        <value-el>gemissement</value-el>
        <value-el>autre</value-el>
</attribute>

<attribute name="type" valuetype="String" />
<attribute name="type classement" valuetype="String" />
<attribute name="flux d air" defaultvalue="non lieu">
        <value-el>non lieu</value-el>
        <value-el>ingressif</value-el>
        <value-el>egressif</value-el>
</attribute>
<attribute name="type de flux" defaultvalue="libre">
        <value-el>libre</value-el>
        <value-el>bloque</value-el>
        <value-el>gene</value-el>
</attribute>
<attribute name="intensite" defaultvalue="non lieu">
        <value-el>non lieu</value-el>
        <value-el>abrupte / intense</value-el>
        <value-el>laxe</value-el>
</attribute>
```

```
        <attribute name="Controle volontaire ou involontaire" valuetype="Boolean" defaultvalue="false" />
        <attribute name="lieu" defaultvalue="non lieu">
                <value-el>non lieu</value-el>
                <value-el>bilabial</value-el>
                <value-el>labio-dental/alveolaire</value-el>
                <value-el>alveolaire</value-el>
                <value-el>palatal</value-el>
                <value-el>velaire</value-el>
                <value-el>glottal</value-el>
        </attribute>
        <attribute name="mouille" valuetype="Boolean" defaultvalue="false" />
        <attribute name="par le nez" valuetype="Boolean" defaultvalue="false" />
        <attribute name="qualite voix" defaultvalue="modal">
                <value-el>modal</value-el>
                <value-el>soupire</value-el>
                <value-el>chuchote</value-el>
                <value-el>murmure</value-el>
                <value-el>creaky</value-el>
                <value-el>tremblante</value-el>
        </attribute>
        <attribute name="voisement" valuetype="Boolean" defaultvalue="false" />
        <attribute name="TP" valuetype="TP" defaultvalue="0"/>
        <attribute name="visible" valuetype="Boolean" defaultvalue="false" />
        <attribute name="peu audible" valuetype="Boolean" defaultvalue="false" />
  </track-spec>

<track-spec name="Forme particuliere" type="primary">
    <attribute name="type" valuetype="String" />
</track-spec>
</body>
</annotation-spec>
```

*Annexe 9 : Nombre d'occurrences de bruits de bouche par type, en fonction de leur flux d'air (opposition bloqué / gêné / continu pour le tableau du haut ; ingressif / égressif pour celui du en bas)*

| Type de bruits de bouche | bloqué | gêné | continu | Total |
|---|---|---|---|---|
| déglutit | 219 | | | **219** |
| click | 63 | | | **63** |
| exp | 6 | 8 | 372 | **386** |
| exp brutale | 22 | 48 | 56 | **126** |
| friction | | 12 | | **12** |
| gemissement | | | 8 | **8** |
| insp | 44 | 43 | 455 | **542** |
| langue-levres mouillees | 79 | | 7 | **86** |
| occlusion | 134 | | | **134** |
| occlusion ingressive_insp | 227 | | 2 | **229** |
| raclement de gorge | | 10 | | **10** |
| relache sa respiration | 109 | | | **109** |
| relachement articulateur | 434 | | | **434** |
| **Total** | **1337** | **121** | **900** | **2358** |

| Type de bruits de bouche | égressif | ingressif | non lieu | Total |
|---|---|---|---|---|
| déglutit | 26 | 26 | 167 | **219** |
| click | | 63 | | **63** |
| expiration | 386 | | | **386** |
| expiration brutale | 126 | | | **126** |
| friction | 3 | 9 | | **12** |
| gémissement | 8 | | | **8** |
| inspiration | | 542 | | **542** |
| langue-lèvres mouillées | 1 | 13 | 72 | **86** |
| occlusion | 74 | 60 | | **134** |
| occlusion ingressive_insp | | 229 | | **229** |
| raclement de gorge | 10 | | | **10** |
| relâche sa respiration | 109 | | | **109** |
| relâchement d'articulateur | 2 | 432 | | **434** |
| **Total** | **745** | **1374** | **239** | **2358** |

## *Annexe 10 : Tableau récapitulatif des données temporelles pour chacun des sujets*

**Données temporelles en secondes selon les tâches liées aux TP**

| Tour de Parole | | Total "0" | "1" | | | Total "1" | Total |
|---|---|---|---|---|---|---|---|
| | Tâches --> | | pendant commentaires | pendant page suivante | pendant prononciation / réponse | | |
| Sujets | F_S | 1657,36 | 374,8 | 40,28 | 111,64 | 526,72 | 2184,08 |
| | F_T | 1903 | 339,76 | 41,52 | 92,96 | 474,24 | 2377,24 |
| | F_M | 1792,72 | 342,24 | 60,24 | 121,6 | 524,08 | 2316,8 |
| | M_J | 1462,84 | 43,88 | 63,24 | 199,24 | 306,36 | 1769,2 |
| | M_R | 1636,28 | 322,08 | 70,48 | 149,44 | 542 | 2178,28 |
| | M_N | 2117,92 | 259,52 | 74,28 | 228,72 | 562,52 | 2680,44 |

## *Annexe 11 : Nombre d'occurrences de bruits de bouche selon leur voisement et leur « qualité de son » en fonction de leur type articulatoire*

| Type de flux d'air du bruit de bouche --> | Qualité de voix | avale salive | click | expiration | expiration brutale | friction | gémissement | inspiration | langue-lèvres mouillées | occlusion | occlusion ingressive_insp | raclement de gorge | relâche sa respiration | relâchement d'articulateur | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Non voisés | chuchote | | | 1 | | | | | | | | | | 1 | 2 |
| | creaky | | | 4 | | | | | | | | | | | 4 |
| | modal | 218 | 61 | 326 | 83 | 12 | | 522 | 85 | 124 | 229 | 8 | 98 | 435 | 2201 |
| | soupire | | | 22 | | | | | 1 | | | | 2 | | 25 |
| | tremblante | | | | | | | 4 | | | | | | | 4 |
| **Total des non voisés** | | **218** | **61** | **353** | **83** | **12** | | **526** | **86** | **124** | **229** | **8** | **100** | **436** | **2236** |
| Voisés | chuchote | | | 3 | | | 1 | | | 1 | | | 1 | | 6 |
| | creaky | | | 1 | | | | 2 | | | | | | | 3 |
| | modal | | | 24 | 43 | | 7 | 14 | | 9 | | 2 | 8 | | 107 |
| | murmure | | | 1 | | | | | | | | | | | 1 |
| | soupire | | | 4 | | | | | | | | | 1 | | 5 |
| **Total des voisés** | | | | **33** | **43** | | **8** | **16** | | **10** | | **2** | **10** | | **122** |
| **Total** | | **218** | **61** | **386** | **126** | **12** | **8** | **542** | **86** | **134** | **229** | **10** | **110** | **436** | **2358** |