Expressions hors des tours de parole : éthogrammes du « *feeling of thinking* »

Fanny Loyau, Véronique Aubergé & Anne Vanpé

Institut de la Communication Parlée
CNRS UMR 5009, Grenoble, France
{loyau, auberge}@icp.inpg.fr

ABSTRACT

During our collect of an expressive corpus, a large quantity of non verbal information has been registered too: top body and face movements, and voice events. We are particularly interested by only these actions which happen outside the talk turn, when the subject thinks, and feels about what he thinks. We want to know if these events are real indices of signals about the mental states or the affective states of the subjects. For that, a typical ethogram methodology has been applied to label these non speech parts into primitive icons of top body movements, face movements and voice events, in order not to take any decision about the interpretation of what could be expressed by these events, but to classify variant movements into minimal icons.

1. INTRODUCTION

Dans le domaine de l'interaction verbale et de la communication expressive, des études de plus en plus nombreuses sont consacrées non seulement aux expressions transmises par le locuteur pendant son tour de parole, mais également aux informations émises par le sujet humain en dehors de son tour de parole, en particulier quand le sujet suit en ligne l'interlocuteur qui est en train de dérouler son tour de parole (« feedback », [3][5]). Il peut alors lui renvoyer des informations sur son attention, l'état de son traitement mental – sa compréhension – sur de ce qu'il reçoit de l'interlocuteur, ses opinions sur ce qu'il reçoit, les émotions que ces traitements induisent sur lui. En dehors de son tour de parole, le sujet peut également être dans une situation de traitement d'une tâche cognitive et /ou physique à accomplir, et faire ou laisser apparaître des informations sur ses états mentaux et affectifs dans le traitement de cette tâche. Cette situation apparaît en particulier fréquemment dans les interactions personne-machine.

Cet article présente les analyses préliminaires des expressions muli-modales d'un corpus expressif multi-locuteurs (Sound Teacher de E-Wiz [1]) dans les parties de l'interaction où les sujets ne sont pas dans leur tour de parole, et dans lesquels cependant les expressions dans la voix, la face ou le corps sont nombreuses et variées. Nous proposons ici les grandes lignes d'une méthode d'annotations de ces expressions qui n'est pas basée sur une mesure automatique de l'image ou du signal vocal, mais qui restreint le rôle de l'expert humain à la détection d'icônes gestuelles ou vocales minimales, sans interprétation de contenu informatif, rejoignant en cela une démarche classique d'éthogramme.

2. LE CORPUS EXPRESSIF SOUNDTEACHER /EWIZ

Le corpus expressif Sound Teacher de E-Wiz [1] a été réalisé à partir de l'enregistrement de 17 sujets, 11 femmes et 6 hommes, placés dans une situation d'apprentissage des voyelles des langues du monde à l'aide d'un pseudo système révolutionnaire, Sound Teacher. Ces sujets sont « piégés » par un scénario de type magicien d'Oz : le sujet pensait communiquer avec un ordinateur, alors qu'en fait le comportement apparent de l'application est géré à distance par le magicien..

Le scénario se déroule en trois grandes phases, la première, dite d'entraînement, familiarise et rassure le sujet, une deuxième phase implique le sujet dans des tâches très simples sur lesquelles il est félicité, ce qui a induit chez l'ensemble des sujets des emotions globalement positives, et une troisième phase, de plus en plus complexe, dans laquelle sont renvoyés au sujet des jugements négatifs, qui se termine par une répétition de la tâche initiale simple, mais en retournant aux sujets de (faux) résultats très mauvais qui les ont soit fortement inquiété, soit déstabilisé. Après chaque enregistrement, les sujets ont auto-annoté leur production en notant selon leur propre choix (langage, dessins, signes etc) leurs états mentaux et affectifs finement au fil de l'avancement du scénario.Les sujets interagissent seulement par la parole, pour les réponses ou pour les phases de commentaires libres (pas de clavier ni souris). Ils sont en isolés en chambre sourde face à un écran, et ne se savent pas enregistrés. La machine dialogue soit par du texte, soit par l'exécution de la demande de tâche. Le sujets sont donc alternativement en phase de lecture, réflexion, production de parole par proposition verbalisée de réalisation de la tâche (sous forme d'un mono-mot mono-syllabique).

3. ETIQUETAGE DU CORPUS : UN ÉTHOGRAMME

3.1. Du "feeling of knowing" au "feeling of thinking"

Sound Teacher est une situation de dialogue minimale, puisque le sujet sait que ses tours de parole ne change pas la nature de l'interaction. La phase de communication humaine ou humanisée dans laquelle le sujet auditeur envoie un feedback à son interlocuteur « intentionnellement » n'est donc pas attendue. Pourtant, nous le montrons plus loin, pendant le « tour de parole » de la machine (lecture), le sujet affiche des expressions

riches pendant la dynamique de la lecture. Surtout, pendant la phase de préparation de sa proposition verbale, les sujets expriment à la fois des affects et des états mentaux. Dans une tâche encore plus spécifique celle de Sound Teacher (un sujet se voit poser une question de culture générale et n'arrive pas à fournir la bonne réponse ; le sujet sait pourtant qu'il connaît cette réponse, il l'a stockée dans sa mémoire, et pourra la retrouver plus tard, mais dans l'instant présent elle n'est pas disponible) des expressions révélant le processus mnésique du sujet ont été observées, étudiées et regrougées comme *feeling of knowing* [6]. La tâche Sound Teacher révélant des processus cognitifs et affectifs plus larges que la tâche mnésique exprimée en *feeling of knowing*, ils seront regroupés ici dans une phénoménologie plus générique que nous appelerons « *feeling of thinking* », expressions des états affectifs et mentaux.

3.2. Méthodologie

Le problèle crucial posé dans cette étude est celui de l'annotation des expressions. Le scenario étant connu, les « états mentaux et affectifs » étant étiquetés par les sujets (et vérifiés pour certains dans des expériences perceptives), la subjectivité de l'étiquetage par un « expert » humain est d'autant plus grande. Nous avons donc fait le choix que les experts (deux experts pour 17 sujets) n'aient pas connaissance a priori des annotations des sujets. Le but est d'utiliser leurs compétences d'humains communicants pour dégager une icônicité minimale des signaux, mais en minimisant leur compétence interprétative (ils ne doivent pas être un participant humain ajouté à l'interaction, mais conserver une distance « objectivante »). Ils doivent se ramener le plus possible à un étiquetage de la « syntaxe icônique » des mouvements et événements vocaux, sans interprétation « sémantique » de l'expression (par exemple, pas d'étiquetage des gestes faciaux en sourire ou autres moues, mais icônes de géometrie et dynamique jugées différentes). Cette démarche qui s'ancre dans une méthodologie d'éthogramme, est donc fondamentalment déterminée par les icînes minimaux définis comme étant les étiquettes à poser sur le corpus. Une difficulté supplémentaire est introduite par la non généricité de certaines icônes que nous avons été amnés à décrire, sans que nous puissions a priori décider si il s'agit d'une variante d'une icône générique (i.e. partagée par tous les sujets, susceptible d'être un signal communicatif) ou idiosynchrasique (i.e. spécifique à un sujet mais néanmoins indice « récupéré » de communication).

Une démarche éthologique

Pour ce faire, nous avons appliquer un protocole issu de l'éthologie (étude des mœurs et du comportement individuel et social des animaux domestiques et sauvages) : nous avons choisi d'annoter nos corpus à l'aide d'éthogrammes. Un tel objet représente l'inventaire des comportements d'une espèce. Plus précisément, « l'éthogramme 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 données 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». Ainsi, nous pouvons étiqueter nos parties des corpus sans parole en utilisant des icônes primitives pour les mouvements du haut du corps, ceux de la face, et les événements vocaux, sans avoir à prendre de décision qui serait du niveau de l'interprétation. Par exemple, le mouvement de la bouche présenté dans la figure 1 ne sera pas étiqueté en tant que « sourire » mais juste en tant qu'icône : « monter le coin des lèvres », que l'on appelle IGS, avec comme variables l'intensité, la durée et l'ouverture ou non de la bouche.

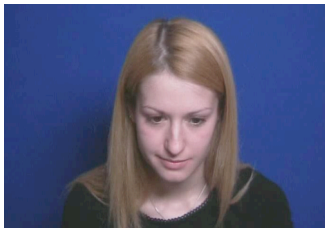

Figure 1 : IGS « monte le coin des lèvres - dissymétrie droite / faible / rapide / fermée ».

Dans la figure 2, se trouve un extrait d'une durée d'environ 25 secondes de l'étiquetage d'un des corpus, avec la partie de l'éthogramme correspondant, où se trouvent les descriptions de chaque icône utilisée dans ce bout d'étiquetage. Chaque occurrence d'icône est numérotée, pour pouvoir ensuite faire des analyses quant à la fréquence d'apparition de chaque icône.

- « eu » (5mn15) : dit « jaune » à 5mn17, IGG11 rapide
- « u » (5mn22) : dit « jaune » à 5mn23 IGF5 tenu
- « é » (5mn25) : dit « vert » à 5mn26 IGF6 tenu
- « o » (5mn30) : IGT3 droite / assez net / rapide, dit « jaune » à 5mn31 puis IGRe1 dans la diagonale, en bas à droite / rapide avec IGSo2 rapide
- « e » (5mn35) : IGLP5 à droite sur ce « e », dit « jaune » à 5mn36 avec IGS3, bouche quasi fermée / long puis IGSo4 rapide pendant lecture consignes

IGF : front plissé
IGLP : lèvres pincées
IGRe : regard hors de l'écran
IGS : coins de la bouche relevés
IGSo : sourcils haussés
IGT : tête penchée sur le coté

IGG : regroupement de plusieurs IG

Figure 2 : Partie de l'étiquetage et son éthogramme associé.

3.3. Signaux vs. indices

Certaines icônes vont sembler se retrouver chez tous les sujets, ce serait donc des signaux, tandis ce que d'autres icônes semblent propres à un sujet, on parlerait d'indices. Mais aucune décision n'est prise a priori sur les différents événements (gestes, face, voix), qui seront plus tard identifiés comme étant soit des signaux de communication, soit des indices biologiques, idiosyncrasiques dont la variabilité est associable à des changements d'états affectifs [2].

4. PREMIERS RÉSULTATS

Les micro et macro organisations temporelles sont fondamentales, soit quand elles sont cohérentes avec l'évolution des états affectifs des sujets, soit parce qu'elles sont révélatrices de ces états (certaines icônes ne seront pas directement porteuses d'information, mais l'organisation temporelle de ces icônes le sera [2]).

4.1. Organisation temporelle

En moyenne, chaque sujet a donné lieu à un corpus d'environ 40 minutes (figure 3).

Le temps alloué aux moments de parole est lui d'approximativement 8 minutes, il y a donc 80% du temps de communication qui se situe hors des tours de parole.

Durant les 32 minutes utilisées par le sujet pour lire les consignes (8 minutes) et surtout pour penser aux réponses qu'il devra ensuite oraliser (24 minutes), nous n'avons pas trouver de position « neutre » pour le haut du corps, le visage, ni de moment totalement silencieux.

Il se passe toujours quelque chose, et c'est ce que nous avons essayé d'étiqueter, de la façon la plus objective possible.

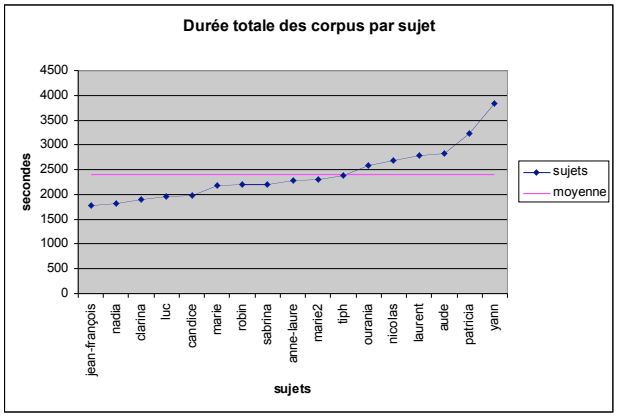

Figure 3 : Durée moyenne des 17 corpus.

Les caractéristiques générales comme les temps de réponse ont été traitées pour tous les sujets, mais pour l'instant l'analyse détaillée présentée ensuite n'a été menée que pour 5 d'entres eux.

Temps de réponse

Le temps moyen de réponse, correspondant au temps entre le moment où le sujet entend les stimuli et celui où il parle pour donner sa réponse est de 4,5 secondes. Cette durée est plus importante pour la phase d'entraînement, diminue dans la phase suivante, positive, puis se rallonge à nouveau dans la dernière phase, regroupant induction négative et déstabilisation (figure 4).

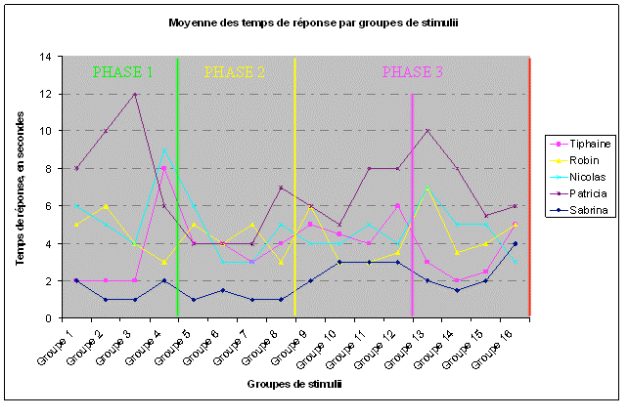

Figure 4 : Moyenne des temps de réponse, par phases.

L'écart type est nettement plus important pour les phases d'entraînement et négative que pour la phase positive, assez stable.

4.2. Expressions du « backchannel »

Nous nous intéressons tout particulièrement au lien qu'il va pouvoir y avoir entre les occurrences des différents indices apparaissant hors des tours de parole et à la fois l'auto annotation faite par les sujets eux-mêmes et les phases du scénario.

Commun aux sujets d'E-Wiz

Voici des mouvements que l'on retrouve chez tous les sujets d'E-Wiz : IGF « plisser le front », IGSo « froncer les sourcils », IGSoh « hausser les sourcils », IGY« plisser les yeux », IGRe « regards hors de l'écran », IGN « plisser le nez », IGLp « passer ses lèvres l'une sur l'autre », IGLm « mordre sa lèvre », IGM « plisser le menton ».

Situation négative

Les regards hors de l'écran : Les sujets regardent tous parfois hors de l'écran, dans toutes les phases, mais surtout lorsqu'ils préparent leur réponse, et plus souvent là encore dans la phase négative. Les moments où ont lieu ces regards sont étiquetés par les sujets eux-mêmes avec en majorité les termes suivants : « perplexité », « doute », « stress », « ennui », « approximation »,« perplexité », « incompréhension », « agacement »… Lorsqu'on s'intéresse aux regards qui ont lue pendant les tours de parole, cela corrobore ces résultats : il y en a plus lors des dernières phases, on retrouve les mêmes étiquettes négatives données par les sujets.

Les rires : contrairement à ce qu'on aurait pu croire a priori, les rires apparaissent plus en situation négative qu'en situation positive. Ils correspondent alors à différentes étiquettes comme « fatigué et amusé », « irrité, anxieux », « stressé, je ne comprends pas » « surpris, nerveux » « doute, très agacé, ri de ma mauvaise performance », « rire = tentative de décontraction », « déception mais m'en amuse », « au pif, une envie de rigoler ». Le rire semble ici avoir pour fonction de permettre au locuteur de changer d'état, de ne pas rester dans une situation négative désagréable. La majorité des

rires ont toutefois lieu pendant les tours de parole, mais aussi lors de situations négatives.

La protrusion des lèvres : cet indice, s'il est présent dans toutes les phases, l'est, lui aussi, plus dans la troisième. Il correspond également à des auto annotations négatives : « j'ai l'air déçue », « concentré », « agacé », « concentration, ennui ». Il n'y en a pas du tout lors des tours de parole.

Les bruits de bouche : Ces bruits, comme des sifflements, des fricatives, des plosions, sont plus nombreux et plus irréguliers dans la phase négative, et augmentent tout particulièrement dans la partie de déstabilisation de cette dernière en devenant également de plus en plus irréguliers. Très peu ont lieu lors des tours de parole.

Ces comportements sont cohérents avec les résultats plus généraux concernant les indices biologiques du comportement observés chez les joueurs de tennis lors de situations inconfortables, de désarroi [2].

Situation positive

Il semble, pour l'instant, que très peu d'indices soient propres à cette situation. Pourtant, les deux premières phases, suivant le scénario, sont censées être positives. Mais dans l'auto annotation faite par les sujets, peu d'étiquettes sont positives, même dans ces deux phases. Les termes les plus fréquents sont les suivants : « ennui », « agacement », « doute »

Le terme de « concentration » est lui aussi utilisé souvent, par tous les sujets, mais selon les autres mots auxquels il est relié il aura une connotation positive (dans les premières phases), ou négative (dans les dernières) : « plus de sérieux, de concentration », « grande concentration le but étant de comprendre ce qui est prononcé » / « ennui, concentration », « concentration – agacement ».

Les sourires : ils sont plus fréquents dans les deux premières phases, et concordent aux étiquettes suivantes : « calme », « assez calme », « concentrée » « fier, content », « étonné et doute », « très fier, content, étonné ». Ils sont donc utilisés à des moments très différents de ceux des rires. Pendant les tours de parole, cela diffère fortement d'un sujet à un autre : chez un il y a en a beaucoup pendant les premières phases, chez un autre il y en a plus qu'hors des tours des tours de parole, et ce pour chaque phase, chez deux autres sujets il n'y en a quasiment pas, 2 dans la deuxième seulement pour l'un de ces sujets, enfin chez le cinquième sujet les sourires pendant le tour de parole ne sont pas majoritaires mais ont lieu principalement lors des dernières phases.

Spécifique au sujet

En ce qui concerne les événements observé chez un seul sujet, il y a l'icône « penche la tête de coté », propre donc à un seul sujet, qui fait ce mouvement principalement avant de répondre à un stimulus sauf dans la sous partie de déstabilisation, et très majoritairement sur sa droite (figure 5).

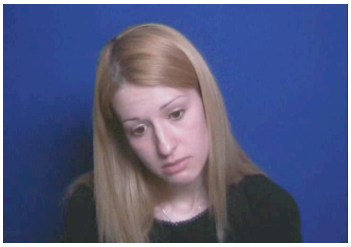

Figure 5 : icône IGT « tête penchée côté droit ».

Cette icône, qui apparaît dans toutes les phases, figure dans des situations positives, étiquetées par le sujet comme « calme », « assez calme », « concentration », mais surtout dans des moments négatifs, étiquetés comme « mal à l'aise », « oppressée », « inquiétude »…Un autre sujet est le seul à renifler, surtout avant de donner sa réponse, et uniquement dans la phase d'entraînement. Dans l'auto annotation cela correspond à des parties où ce sujet dit être « concentrée », ou « concentrée, stressée ».

5. CONCLUSION

Dans ce travail très préliminaire, nous avons essayé d'ébaucher une méthodologie encore empirique d'éthogrammes des comportements expressifs des sujets dans l'interaction mais hors de leur tour de parole. Nous avons déjà pu observer que l'organisation temporelle joue un rôle fondamental, aussi bien au niveau global que local. Nous allons confronter les icônes que nous avons identifiées d'abord à une validation statistique de leur pertinence, puis à leur validation perceptive, à la fois dans des tests de perception d'icônes isolées, ou de schémas rythmiques isolés, et en synthèse avec l'agent conversationnel, « GRETA » [4].

BIBLIOGRAPHIE

[1] V. Aubergé, N. Audibert, and A. Rilliard. E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab. *4th LREC*, 179-182, 2004.

[2] G, Carlier, and C. Graff, to be published. Unpredictability as a counter strategy: An analysis of elite matches. *Journal of Sciences*, 2006.

[3] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi. A model of attention and interest using gaze behavior. *IVA'05 International Working Conference on Intelligent Virtual Agents*, 2005.

[4] I. Poggi, C. Pelachaud, F. de Rosis, V. Caroglio, B. de Carolis. GRETA. A Believable Embodied Conversational Agent. *Multimodal Intelligent Information Presentation*, O. Stock and M. Zancarano, eds, , Kluwer, to appear, 2005.

[5] M. Schröder, D. Heylen and I. Poggi, to be published. Perception of non-verbal emotional listener feedback. *Speech Prosody 2006*.

[6] M. Swerts and E. Khramer. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 2004.