HAL
open science

A la recherche d'indices de culture et/ou de langue dans les micro-événements audio-visuels de l'interaction face à face

Rosario Signorello, Véronique Aubergé, Anne Vanpé, Lionel Granjon, Nicolas Audibert

To cite this version:

Rosario Signorello, Véronique Aubergé, Anne Vanpé, Lionel Granjon, Nicolas Audibert. A la recherche d'indices de culture et/ou de langue dans les micro-événements audio-visuels de l'interaction face à face. WACA 2010 - Workshop Francophone sur les Agents Conversationnels Animés, Nov 2010, Lille, France. pp.69-76. hal-00542004

HAL Id: hal-00542004

https://hal.science/hal-00542004

Submitted on 18 Jan 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A la recherche d'indices de culture et/ou de langue dans les micro-événements audio-visuels de l'interaction face à face

Rosario Signorello[1] Véronique Aubergé[1,2] Anne Vanpé[1,2] Lionel Granjon[1] Nicolas Audibert[3]
{rosario.signorello, veronique.auberge, anne.vanpe, lionel.granjon}
@gipsa-lab.grenoble-inp.fr ; nicolas.audibert@gmail.com

[1] GIPSA-Lab, Département Parole et Cognition, UMR CNRS 5216, Université Stendhal, Grenoble, France

[2] Laboratoire d'Informatique de Grenoble, équipe GETALP, UMR CNRS 5217, Grenoble, France

[3] Laboratoire Informatique d'Avignon, Université d'Avignon, Avignon, France

Résumé :
Cet article présente une étude sur l'identification perceptive, par des juges français et italiens, de la langue/culture, le degré de confiance accordé à ce choix, et, en parallèle, du degré d'extraversion. Les jugements sont faits sur des micro-événements acoustiques et visuels non langagiers (ex. bruits de bouche, *grunts*, *bursts*, *fillers* et interjections) de 6 sujets français, placés en interaction personne-machine. Ces événements ont été choisis et ordonnés en listes de 24 stimuli selon un contrôle croissant de la prosodie (à partir d'un non contrôle) et de leur distance des sons phonologiques. Pour chaque modalité de passation (Audio seul ou Visuel seul ou Audio-Visuel), des groupes différents de 15 juges, de chaque nationalité, ont été interrogés. Les résultats montrent une tendance, des juges des deux nationalités, à identifier les sujets comme français. 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 le choix langue/culture et du degré de confiance.
Mots-clés : culture, langue, extraversion, contrôle prosodique, micro-gestualité, interjections, *bursts*, *grunts*.
Keywords: culture, language, extraversion, prosody, micro-gestures, interjections, bursts, grunts.

1 Introduction

Beaucoup de travaux [ex. 2] ont été consacrés aux relations étroites entre langue et culture, aussi bien dans leurs aspects théoriques que dans des champs applicatifs comme l'enseignement des langues destiné à un usage pertinent en situation socio-culturelle. Les sciences du langage, mais aussi tous les travaux d'ingénierie qui s'adressent à la mise en situation réelle d'agents virtuels ou de robots humanoïdes, pointent sur un niveau de granularité révélant des structures et des modèles caractéristiques du « rôle social » tenu dans chaque contexte situationnel : un banquier, un enseignant, un assistant, une mère s'adressant à son bébé, etc. Bien entendu la description de ces rôles (plus ou moins emprunts d'invariants) se construit par le filtre de la langue/de la culture dans lesquelles ils se définissent.

L'interaction est un processus dynamique contextualisé par quand, où, pourquoi, à qui et qui est en cours de communication langagière [1]. Lorsque l'on observe une interaction en train de se dérouler les caractéristiques comportementales de la personne (ses tendances génériques et ses spécificités de personnalité, de langue et de culture) deviennent certainement des paramètres fondamentaux pour décrire, comprendre, simuler et prédire ses actes de communication dans cette entité globale de l'interaction. L'interactant est en communication permanente, pendant ses tours de parole (bien sur) mais aussi en dehors ceux-ci, par des gestes et micro-gestes audibles et/ou visibles qui se réalisent dans ce que nous désignons comme *Feeling of Thinking* (FoT) : états mentaux (ex. la compréhension, la concentration), attitudes, affects sociaux (ex. le doute, la politesse), émotions (ex. la tristesse) et humeurs (ex. le stress) [3, 8, 9].

2 Objectifs de l'étude

Les comportements communicatifs d'un humain, en interaction face à face, mettent en

jeu autant ses composantes personnelles biophysiques et mentales (ex. sa personnalité) que le rôle sociétal qui situe son interaction (cela avec le filtre prégnant de sa langue et de sa culture). Il est supposé ici que l'observation des événements langagiers d'une personne, au cours d'une interaction, ne pourra être expliquée globalement (et donc à terme prédite ou reconnue automatiquement par des agents virtuels, par exemple) que si sont pris en compte tous ces degrés de granularités emboités : sa personnalité, son rôle sociétal au moment de l'interaction, sa langue/sa culture.

L'expérience perceptive qui est présentée ici est une étude très préliminaire qui se donne comme objectif de poser deux questions fondamentales :

A. Peut-on discriminer deux langues/cultures, mêmes proches, à travers des micro-événements, audibles et visibles, mais non langagiers (bruits de bouche, *grunts, bursts*, *fillers*, interjections) du FoT ? ;
B. Existe-t-il une interaction perceptive entre des faits de langue/culture et des indices de personnalité ?

Le problème posé ici est lié à la nature de ces micro-événements non langagiers : sont-ils de valeur informative variée par rapport aux critères d'identification de la langue/culture ? L'hypothèse forte est que la nature langagière des événements sonores est construite par le contrôle de la prosodie (i.e. contrôle « volontaire » de la durée, de la qualité et/ou de la rythmicité), et que les premiers faits de langage apparaissent ainsi bien avant que ces sons ne soient doublement articulés (controverse introduite par [8], inspirée par [7]). Nous voulons donc tenter 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. 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). Enfin, des stéréotypes de représentations collectives (ex. le trait d'une supposée extraversion dans l'interaction langagière des habitants proches de la méditerranée) sont bien installés et véhiculés par la croyance populaire.

3 Protocole expérimental

3.1 Participants

90 juges[1], 45 français et 45 italiens, ont participé au test de perception. Chaque groupe comptait 3 sous-groupes de 15 pour chaque modalité de passation du test : Audio seul (noté A), Visuel seul (noté V) et Audio-Visuel (noté AV). Chaque juge a passé le test dans une seule modalité. La faible connaissance de l'autre langue/culture était une condition nécessaire (ex. ne pas être bilingue français-italien et/ou avoir vécu dans les deux pays).

3.2 Choix et organisation des stimuli

Les micro-événements étudiés ici sont extraits du corpus d'expressions émotionnelles authentiques Sound Teacher recueilli sur des sujets français grâce à la plate-forme magicien d'Oz E-Wiz [1], induisant des états affectifs « prédits », dans une tâche prétexte d'apprentissage phonétique des sons des langues du monde. Après l'expérimentation, les sujets impliqués auto-annotaient leurs propres enregistrements, désignant ainsi « naïvement » un large spectre d'états mentaux, cognitifs, attitudinaux et plus largement affectifs, qui a été regroupé en FoT [3]. Un étiquetage des formes acoustiques et gestuelles a été mené ensuite avec une méthodologie empruntée à l'éthologie [9]. Cela a été fait afin de ne pas inscrire, *a priori,* les données observées dans une théorie de la gestualité ou de la face, et de vérifier précisément, *a posteriori*, la capacité de ces

[1] Moyenne d'âge 26,33 ans. Juges français : 35 femmes, 10 hommes ; Juges italiens : 15 femmes, 35 hommes.

modèles à prendre en compte aussi bien les valeurs annotées par le sujet (qui dépassent largement le cadre des émotions) et les formes acoustiques et gestuelles ([3] puis [8]). Ces dernières sont désignées comme « subtiles » par certains auteurs [4, 5] sans pour autant couvrir l'étendue de celle présentes et validées dans Sound Teacher (depuis [1] à [9]).

Les sujets étaient tous français : 3 femmes (F-M, F-S, F-T) et 3 hommes (M-J, M-N, M-R), choisis pour une variabilité supposée (mais non vérifiée objectivement) de leur extraversion. 24 stimuli par sujet ont été sélectionnés, par expertise auditive de leur articulation et par analyse du signal acoustique [8]. Ils ont étés ensuite ordonnés selon leur qualité prosodique croissante, cette prosodie étant supposée informative. Dès sons non phonétiques (supposés sans contrôle prosodique volontaire, ex. bruits de bouches), puis, avec un contrôle de plus en plus précis (faisant apparaître des sons non phonétiques avec contrôle prosodique, ex. expiration, qualité de voix) jusqu'aux phonèmes de la langue qui n'entrent pas dans le processus du morpho-lexique (ex. *fillers*, interjections). Ci-dessous en détail, du 1er au 24ème stimulus :

- 1er stimulus → 10ème/14ème stimulus[2] : pas de contrôle prosodique volontaire des articulateurs (relâchement articulateur, respiration, avalement de la salive, interaction langue-lèvres, friction, aspiration salive) ;
- 10ème/14ème stimulus → 13ème/20ème stimulus : inspiration ou expiration contrôlée dans le temps, raclement de gorge contrôlé au niveau de leur durée, clic ou occlusions non phonétiques soit répétés soit suivis d'une expiration ou d'une inspiration, contrôle de qualité de voix dans le temps, etc. ;
- 13ème/20ème stimulus → 22ème/24ème stimulus : *fillers* (ex. « *mmmh* », « *euh* », « *fff* »), sons phonétiques (voyelles, syllabes) qui peuvent être ou non allophones du français ;
- 22ème/24ème stimulus → 24ème stimulus : interjections (ex. « *ben* », « *ouh la la* », etc.), sons phonologiques du français.

[2] Si deux valeurs sont utilisés (ex. 10ème/14ème stimulus) c'est parce qu'il n'a pas été possible de trouver exactement des stimuli aux caractéristiques articulatoires et acoustiques similaires pour chaque sujet. Toutefois, l'ordre supposé croissant a été respecté pour chaque sujet.

3.3 Choix des modalités de présentation

Un point clé de cette étude a été de vérifier si les informations traitées visuellement (les traits statiques d'aspect vestimentaire, capillaire, d'accessoires ainsi que les traits dynamiques de gestualité, liés soit à la personnalité soit à des facteurs culturels et langagiers de communication) sont en rapport avec celles traitées acoustiquement (en particulier la dynamique de la prosodie testée ici). Il était donc fondamental que les facteurs d'apprentissage dans une modalité n'influencent pas une autre modalité. C'est pourquoi chaque modalité (A, V ou AV) a été jugée par des groupes différents (de 15 juges chacun), et ceci pour chacune des deux nationalités d'appartenance.

Pour vérifier, en particulier, que les signaux acoustiques (modalité A) contiennent, à partir d'un certain seuil, des indices informationnels propres aux formes acoustiques, il a été nécessaire de présenter les mêmes stimuli, dans le même ordre, en modalité V. La modalité AV, quant à elle, a permis de mesurer comment sont intégrées ces informations.

3.4 Procédure

L'interface du test affichait, dans sa première page, les instructions détaillées avec les consignes pour le bon déroulement du test. Les 24 stimuli de chaque sujet étaient présentés aux juges dans l'une des modalités possibles de présentation des stimuli (choisie au hasard par l'expérimentateur). L'ordre d'affichage des sujets était aléatoire et différent pour chaque juge. Après la présentation de chaque stimulus, un écran de réponse s'affichait :

- une première échelle de réponse permettait aux juges de choisir la langue/culture (soit français, soit italien) et donner en même temps la valeur (entre 1 et 10) de la confiance accordée à leur choix ;

- Les juges déplaçaient ensuite un autre curseur sur un axe limité par « introversion » et « extraversion », sans autre valeur quantitative que le milieu noté « 0 ». L'objectif était ici de mesurer si les stimuli donnaient, ou non, aux juges « matière » à évaluer les sujets sur les traits de personnalité introverti *vs.* extraverti, mais aussi de vérifier si ce jugement était influencé par le jugement de langue/culture et s'il évoluait avec la progression des stimuli.

Dans chaque modalité de passation, afin de mesurer indirectement un potentiel facteur d'apprentissage, un jugement « a posteriori » global pour le sujet locuteur a été demandé aux juges, à la fin de la série de stimuli et sans support de stimulus. Cette mesure leurs a été demandée pour le choix catégoriel italien *vs.* français (avec le degré de confiance associé à cette réponse) et pour le degré d'extraversion perçu.

Ce test a été passé devant un ordinateur (résolution écran 1440x900 pixels) à l'aide d'un casque Sennheiser HD 25-13. Résolution des stimuli : Vidéo 720x540 pixels ; Audio 44.1 KHz, 16 bits.

4 Résultats

4.1 Discrimination

L'un des buts de notre étude était de savoir si les juges arrivaient à percevoir une différence, en terme d'appartenance culturelle, entre les sujets : ont-ils été reconnus comme des français ? Y-a-t'il eu un effet de la nationalité des juges sur ce choix ? La modalité de passation du test a-t-elle influencé la perception du sujet ? Et si oui, de quel sujet en particulier ? Les sujets ont-ils été eux-mêmes source d'influence sur la perception des juges ?

Globalement, pour les deux nationalités de juges (toutes modalités confondues), les sujets sont perçus à 58,4% comme français et à 41,6% comme des italiens ($\chi^2(1)=15.79$, $p<.001$). Les juges français classifient mieux : ils identifient les sujets comme français à 60,2% ($\chi^2(1)=267.26$, $p<.001$) tandis que les juges italiens à 56,7% ($\chi^2(1)=116.80$, $p<.001$). Les réponses « a posteriori » confirment ces tendances : juges français 78,9% ($\chi^2(1)=90.13$, $p<.001$), juges italiens 72,6% ($\chi^2(1)=55.12$, $p<.001$).

4.2 Croisement des facteurs « modalité », « sujet » et « choix de langue/culture »

4.2.1 Modalité A (Audio seul)

Comme le montre le TAB. 1a les juges français, en moyenne, reconnaissent tous les sujets comme des français, significativement mieux que le hasard ($p<.001$, sauf M-R ($p>.05$)). Ce résultat est confirmé par le jugement « a posteriori » ($p<.001$, sauf M-J ($p>.05$)). Au contraire, les juges italiens (TAB. 1b) reconnaissent 4 sujets comme italiens (M-J, M-N, M-R ($p<.001$) et F-S ($p<.05$) ; résultats pour F-M et F-T pas significatifs, $p>.05$). Ces tendances changent radicalement en considérant les moyennes des réponses « a posteriori » : presque tous les sujets sont reconnus comme des français ($p<.001$), mis à part M-J qui continue à être perçu comme italien avec un pourcentage assez élevé (73,3%, $p<.001$).

Grace a ces résultats nous pourrions déjà avancer l'hypothèse qu'il y a de l'information dans l'audio mais qu'elle n'est pas suffisante pour discriminer la langue/culture (au moins pour les juges italiens qui semblent subir un biais cognitif, ex. l'attraction vers leur propre langue/culture d'appartenance).

4.2.2 Modalité V (Visuel seul)

Dans cette modalité, deux sujets (F-M et F-S) sont perçus avec des tendances similaires par les juges français et italiens. Les juges français perçoivent quatre sujets (F-T, M-J, M-N, M-R) comme des français ($p<.001$) et deux sujets (F-M et F-S) comme des italiens ($p<.001$) (cf. TAB. 1a). Les juges italiens reconnaissent deux sujets (M-N et M-R) comme des français ($p<.001$) et deux sujets (F-M ($p<.001$) et F-S ($p<.05$)), comme des italiens. Les résultats

pour les sujets F-T et M-J ne sont pas significatifs (p>.05) (cf. TAB. 1b). Dans les réponses « a posteriori » les juges français aussi bien que les juges italiens perçoivent quatre sujets (F-T, M-J, M-N et M-R) comme des français (p<.001) et le sujet F-M comme italien (p<.001). Le résultat pour le sujet F-S n'est pas significatif (p>.05).

D'après ces résultats similaires nous pourrions avancer une hypothèse sur l'aspect général des locuteurs et/ou des *a priori* stéréotypiques d'aspect (traits somatiques, indices esthétiques, etc.) partagés par les deux cultures testées. Toutefois, il est nécessaire d'analyser l'évolution des réponses durant la progression des listes de stimuli (cf. 4.3), pour savoir si ces résultats doivent être reliés à l'aspect global du sujet ou aux stimuli dynamiques présentés en liste (soit par apprentissage quantitatif, soit par qualité informative – l'hypothèse « prosodique » étant donnée sur l'acoustique mais elle pourrait aussi être visible).

4.2.3 Modalité AV (Audio-Visuel)

Les juges français reconnaissent, en moyenne, trois sujets comme français (F-M, M-N, F-S, p<.001) et un sujet comme italien (M-R, p<.05) (résultats pour F-T et M-J non significatifs, p>.05) (cf. TAB. 1a). Par contre, dans les réponses « a posteriori » les sujets sont tous reconnus comme français (p<.001). Les juges français ne sembleraient pas suivre les performances de la modalité A. Nous interprétons cela par le fait que l'audio serait porteur d'informations et que les juges français pensent y puiser des indices. Par contre, avec l'ajout de la vidéo, ces indices ne seraient pas suffisamment discriminants pour diriger les juges français vers le bon choix. Les indices visuels auraient donc affecté leur perception.

Au contraire, les juges italiens ne reconnaissent que des français parmi les sujets (cf. TAB. 1b) (les réponses « a posteriori » confirment ces tendances). Ces résultats légitiment donc l'interprétation formulée pour la modalité A : l'information contenue dans l'acoustique (insuffisante si présentée seule), ajoutée à celle contenue dans la visuelle, fait basculer les résultats vers la bonne réponse.

Comme pour la modalité V, ces interprétations doivent être confrontées à l'évolution des performances des juges dans la progression des stimuli (cf. 4.3).

TAB. 1 : Matrices de scores de choix langue/culture en pourcentage par modalité de passation (A, V, AV), choix langue/culture (*fra vs. ita*) et sujet. Analyse statistique : t-tests échantillons uniques, valeurs de test 50%. Niveaux de significativité : *= p<.001 ; ^= p<.05 ; °= choix non significatif.

1a - Juges Français

	A		V		AV	
	fra	ita	fra	ita	fra	ita
F-M	77,2*	22,8	27,8	72,2*	73,3*	26,7
F-S	75*	25	42,8	57,2*	63,1*	36,9
F-T	66,1*	33,9	66,7*	33,3	51,4°	48,6°
M-J	66,4*	33,6	62,2*	37,8	49,7°	50,3°
M-N	68,1*	31,9	73,1*	26,9	70,8*	29,2
M-R	48,6°	51,4°	55,8*	44,2	44,7	55,3^
Total	66,9*	33,1	54,7°	45,3°	58,8*	41,2

1b - Juges Italiens

	A		V		AV	
	fra	ita	fra	ita	fra	ita
F-M	53,3°	46,7°	39,2	60,8*	73,3*	26,7
F-S	44,2	55,8^	44,4	55,6^	68,1*	31,9
F-T	48,9°	51,1	55^	45	66,4*	33,6
M-J	26,9	73,1*	55^	45	59,2*	40,8
M-N	40,3	59,7*	60,6*	39,4	66,1*	33,9
M-R	40,8	59,2*	91,9*	8,1	87,2*	12,8
Total	42,4	57,6*	57,7*	42,3	70,0*	30,0

4.3 Point de stabilité de la réponse : degré de confiance et choix langue/culture

La FIG. 1 montre l'évolution du degré de confiance qu'ont les juges sur leurs réponses en fonction de leur exposition aux stimuli. Le degré de confiance est modérément mais significativement corrélé à l'ordre de présentation des stimuli dans les 3 modalités

de présentation et pour chaque nationalité des juges. Cette corrélation étant toutefois faible en condition V : juges français (A (r=.516), V(r=.095), AV (r=.486) ; corrélations significatives à p<.0001) ; juges italiens (A (r=.289) ; V (r=.103), AV (r=.352), corrélations significatives à p<.0001).

La comparaison, au moyen de tests χ^2, des choix langue/culture effectués par les juges pour différentes séquences de stimuli (définies en découpant l'ensemble des 24 stimuli en 2, 3 ou 4 parties égales) confirme que cette augmentation du degré de confiance s'accompagne d'une évolution significative de la distribution des réponses.

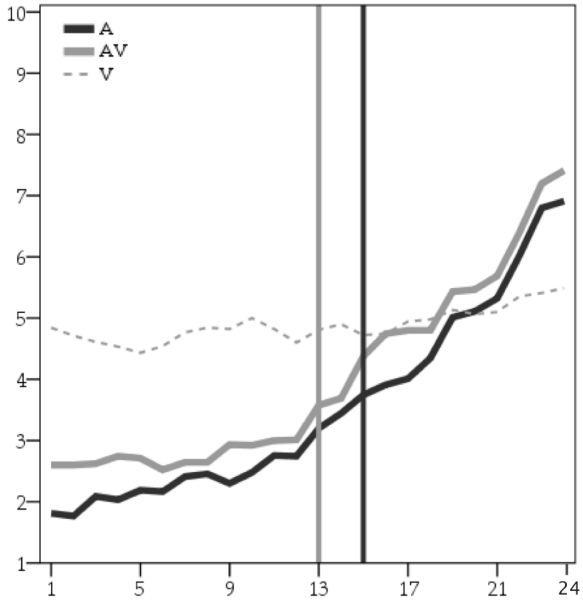
Nous avons voulu ensuite vérifier si une stabilisation des réponses se manifestait à l'apparition des stimuli contrôlés prosodiquement. Nous avons donc extrait pour chaque triplet (juge, sujet, modalité) le « point de stabilité » (cf. TAB. 3), défini comme l'indice du stimulus à partir duquel le juge ne modifie plus son choix de langue/culture et le degré de confiance dans ce choix a une augmentation plus importante par rapport aux stimuli précédents.

TAB. 3 : « Point de stabilité » du choix langue/culture par modalité de passation (A, V, AV), nationalité des juges (FRA *vs*. ITA) et sujet.

	A		**V**		**AV**	
	FRA	ITA	FRA	ITA	FRA	ITA
F-M	11,67	16,40	6,53	12,40	9,80	13,40
F-S	11,33	16,20	8,60	7,33	11,87	8,87
F-T	15,53	15,80	8	10,20	16	11,27
M-J	15,93	10,93	7,80	7,13	16,47	13,87
M-N	18,07	15,13	10,07	10,07	10,20	12,93
M-R	18	19,07	5,73	5,73	17,73	8,93
Total	15,09	15,59	7,79	8,61	13,68	11,54

Comme le montre le TAB. 3 les réponses des juges se stabilisent vers le 15ème stimulus pour la modalité A et un peu plus précocement, autour du 12ème/13ème stimulus, pour la modalité AV. Cette stabilité corresponde à l'apparition des stimuli pré-phonétiques ou phonétiques, contrôlés prosodiquement.

1a – Juges Français



1b – Juges Italiens

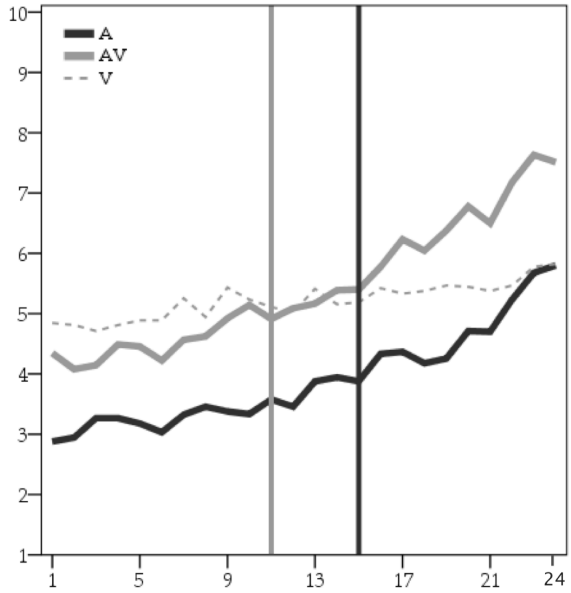


FIG. 1 : évolution du degré de confiance par rapport à l'ordre des stimuli (de 1 à 24) et à la modalité de passation du test (A, V, AV). Lignes verticales : « point de stabilité » (A et V ; non significatif en V).

Par contre, aucun point de stabilité significatif n'a été relevé dans la modalité V. Le choix de langue/culture, ainsi que le degré de confiance, restent stables pendant la présentation de la

suite des stimuli. Chose qui va dans le sens de l'hypothèse prosodique : les informations visuelles seraient ainsi principalement liées à l'aspect global, les micro-événements n'apporteraient pas d'information dynamique.

Le point de stabilité diffère pour tout sujet : comme dit *supra* (cf. 3.2), les stimuli sélectionnés ne correspondent pas en terme de nature et de placement dans la suite d'affichage des stimuli.

4.4 Degré d'extraversion

Des patterns de réponse différents, par rapport à la modalité de passation du test et à la nationalité du juge, se sont avérés pour les sujets : F-M est toujours perçu comme introverti ; F-S est vu comme plus ou moins extraverti ; F-T comme extraverti en A, introverti en AV et V; M-J comme introverti en A, comme extraverti en AV et V; M-N comme extraverti en A et comme introverti en AV et V; enfin, M-R comme extraverti en A et comme introverti en A et V (Analyse statistique : ANOVA à mesures répétées avec comme facteurs *inter-sujet* la modalité de passation du test ($F(8,21)=4.19$, $p<.001$) et la nationalité du juge ($F(4,90)=1.05$, $p=0.382$)). Les réponses « a posteriori » confirment ces tendances.

L'autre facteur qui semble influencer le jugement est l'ordre d'affichage des stimuli. Les juges français, plus que les juges italiens, montrent une tendance à noter comme introverti pendant les premiers stimuli, pour progresser relativement peu vers extraverti. Une analyse sur la régression linéaire de ces progressions montre qu'il existe une pente significative entre le début et la fin des stimuli (juges français ($F(1,64)=136.61$, $p<.001$) ; juges italiens ($F(1,64)=28.65$, $p<.001$)).

Enfin, nous n'avons pas constaté une interaction significative entre degré d'extraversion perçu et choix langue/culture.

5 Conclusions et perspectives

Ce travail préliminaire sur la discrimination de deux langues/cultures très proches, à partir de micro-événements – expressions du FoT – de 6 sujets français, a permis de considérer comme « valide » l'interrogation sur l'informativité langagière/culturelle de ces événements subtils, *a priori* non linguistiques.

La modalité de passation du test est l'un des facteurs qui a influencé le choix des juges. Ainsi, l'acoustique présentée seule ne s'est pas avérée suffisamment informative (en particulier pour les juges italiens), au point de ne pas leur permettre de faire le choix correct. Mais, la progression qualitative de l'information véhiculée et l'accumulation des stimuli contrôlés prosodiquement, ont permis aux juges de corriger leur choix. Les tendances de la modalité AV montrent cependant que (dans le cas des juges italiens) l'audio permet, en s'ajoutant à la vidéo, de faire basculer les résultats vers la bonne réponse. En particulier, dans ces deux modalités, l'ordre des stimuli a révélé un changement – au point de stabilité – dans le choix langue/culture. Cela précisément à l'apparition des stimuli donnés comme premier niveau informationnel selon l'hypothèse prosodique croissante de départ (supposée rapprocher ces éléments de faits de langue). D'autre part, les résultats de la modalité V nous montrent que les scores de reconnaissance des juges sont dus, vraisemblablement, à l'aspect des sujets. En visuel la réponse est précoce et correspond à un degré de confiance élevé. Les juges restent sur leur *a priori* de départ, correct ou non. D'ailleurs, aucun point de stabilité des réponses n'a été relevé, ce qui prouve donc que dès le départ les réponses sont stables (ou ne le sont jamais ce qui n'est pas le cas ici).

Enfin, le facteur d'extraversion, proposé *a priori* pour vérifier si ce trait « populaire stéréotypique » a une réalité tangible, ne s'est pas avéré pertinent. Cependant, il n'est pas possible d'en déduire que ce trait n'a pas eu d'influence sur la discrimination perceptive français *vs.* italien parce que aucune mesure « objective » sur la personnalité des sujets n'avait été menée en amont. Il serait donc intéressant, dans une expérience ultérieure, de

mesurer *a priori* le degré d'extraversion des sujets et de les choisir dans un spectre large sur ce trait.

Pour compléter cette étude, et en tirer des conclusions plus solides, il est nécessaire de croiser symétriquement les juges et les sujets dans chacune des deux langues/cultures. En outre, il faudrait reproduire cette expérience croisée sur des listes progressives de stimuli qui commencent précisément au premier niveau de contrôle prosodique tel qu'il a été validé dans cette expérience.

Ce type d'études des influences réciproques, en perception, des facteurs de langue/culture et de personnalité, pendant les interactions verbales, pointent sur l'importance de tenir compte de ces facteurs si l'on veut donner à un agent virtuel animé un comportement « culturellement » pertinent en situation, réaliste et écologique. Les « détails », non seulement dans la dynamique visuelle, mais également – c'est notre propos – dans la « périphérie » de la parole, commencent certainement à peine à nous montrer l'extrême importance que nous devons porter à sa pertinence cognitive et la nécessité de son réalisme virtuel. De plus, la présence ou l'absence de facteurs de langue/culture et de personnalité peuvent modifier et/ou activer des stratégies spécifiques dans l'agent virtuel pour la planification du discours et pour le choix des signaux de communication à utiliser [6]. Il s'agit de donner une capacité de perception « de haut niveau cognitif » à ces agents. Si l'agent a un petit modèle de l'utilisateur (i.e. s'il réussi à comprendre comme il est vraiment, d'où il vient, quel type de personnalité il a, etc.) il peut choisir, dans sa pré-planification de comment répondre à l'utilisateur, d'interagir d'une façon plutôt que d'une autre. Pour l'agent virtuel est une façon de décider : *« en fonctionne de comme il/elle est, je lui dis ça ou pas »* ?

Références

[1] Aubergé, V., Audibert, N., et 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), pages 499-528.

[2] Lévi-Strauss, C. (1958). *Anthropologie structurale*. Plon, Paris.

[3] Loyau, F., et Aubergé, V. (2006). Expressions outside the talk turn: ethograms of the feeling of thinking. In *5th LREC*, pages 47-50.

[4] Matsumoto, D. (s.d.). Subtle Expression Recognition Training (SubX) [méthode]. URL : http://www.humintell.com/subtle-expression-recognition-training
Dernière visite 17/10/2010.

[5] Matsumoto, D., Consolacion, T., Yamada, H., Suzuki, R., Franklin, B., Paul, S., et al. (2002). American-Japanese cultural differences in judgements of emotional expressions of different intensities. *Cognition & Emotion*, 16(6), 721-747.

[6] Poggi, I., Pelachaud, C., de Rosis, F., Carofiglio, V., et De Carolis, B. (2005). GRETA. A Believable Embodied Conversational Agent. In Stock, O. and Zancarano, M., éditeurs : *Multimodal Intelligent Information Presentation*, pages 1–23. Kluwer, Belgique.

[7] Rossi, M. (1977). L'intonation est la troisième articulation. *Bulletin de la Societé Linguistique de Paris,* LXXII, (1), pages 55-68.

[8] Vanpé, A. et Aubergé, V. (2010a). Expressions des états mentaux/cognitifs et affectifs : Prosodie des productions vocales minimales – des grunt et burst à l'interjection. In *XXVIIIe JEP*.

[9] Vanpé, A. et Aubergé, V. (2010b). Prosodie expressive audio-visuelle de l'interaction personne-machine. *Technique et Science Informatiques, numéro spécial Agents Conversationels Animés* 29.