Emotions actées vs. spontanées : variabilité des compétences perceptives

Nicolas Audibert[1], *Véronique Aubergé*[1] *et Albert Rilliard*[2]

[1] Gipsa-lab, Dépatement Parole & Cognition (Institut de la Communication Parlée),
CNRS UMR 5216/Université Stendhal/INPG, 38040 Grenoble Cedex 9, France
{Nicolas.Audibert, Veronique.Auberge}@gipsa-lab.inpg.fr
[2] LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
Albert.Rilliard@limsi.fr

ABSTRACT

This paper reports the results of a discrimination experiment of acted vs. spontaneous expressive speech by naive listeners. Monoword utterances of 4 French-speaking actors trapped in a Wizard of Oz before simulating in an acting protocol supposed to be optimal for them were extracted from the Sound Teacher/E-Wiz multimodal corpus. Pairs of acted vs. spontaneous stimuli, expressing affective states related to anxiety, irritation and satisfaction were discriminated by 33 French listeners in audio-only (A), visual-only (V) and audiovisual (AV) conditions. 70% of listeners were able to identify acted vs. spontaneous pairs over chance in V, 78% in A and 85% in AV. A strong listener effect confirms the hypothesis of variable competence for separating involuntary vs. simulated affects. Perceived emotional intensity differences appear as a strong cue to discrimination but cannot account for the whole variability.

Keywords: spontaneous vs. acted expressive speech, perceptual discrimination, individual variability

1. INTRODUCTION

Les études portant sur la parole expressive ont connu un très large essor au cours de la dernière décennie, la plupart d'entre elles s'appuyant sur des productions actées. Bien que la question de la validité des corpus expressifs produits par des acteurs ait été largement débatue (voir par exemple [5] pour une discussion détaillée des implications du recours à de tels corpus), peu d'études à notre connaissance se sont focalisées sur les différences en production et/ou en perception entre parole actée et spontanée. Aubergé et al. [3] ont montré que l'amusement acté pouvait être discriminé de l'amusement spontané, avec une grande variabilité dans les performances des juges indépendamment des compétences d'acteur des locuteurs. D'autre part Wilting et al. [12] ont induit des humeurs positives et négatives chez des locuteurs néérlandais sans compétences d'acteurs particulières, qui ont ensuite reproduit les mêmes énoncés en simulant les humeurs ressenties. Une évaluation perceptive a montré que les expressions actées étaient perçues comme plus intenses que les expressions spontanées.

A la suite des travaux de Fonagy [8] et Scherer [11], nous émettons l'hypothèse que les expressions des affects sont d'abord distinguées cognitivement en fonction du mode de contrôle par le locuteur : volontaire vs. involontaire, c'est-à-dire ici acté vs. ressenti et exprimé spontanément, plutôt qu'en fonction de la valeur de l'affect exprimé. Dans cette optique [2], tandis que les émotions authentiques seraient produites par un contrôle involontaire lié à l'effet *push* dans le modèle proposé par Scherer [11], les productions contrôlées volontairement ou affects sociaux que nous désignons sous le terme d'attitudes s'appuieraient sur la compétence de simulation du locuteur via la boucle cognitive de simulation mise en évidence par Damasio [7]. Notre hypothèse est que cette compétence de simulation, centrale dans la communication expressive, est également utilisée par les acteurs, tout particulièrement lorsque ceux-ci s'appuient sur des méthodes comme l'élicitation dans lesquelles l'expression d'affects s'appuie sur le souvenir d'épisodes émotionnels [6], comme c'est le cas des acteurs participant à cette étude.

La principale question à laquelle nous tentons de répondre est donc si des énoncés actés vs. spontanés exprimant des valeurs d'affects similaires peuvent être discriminés par des auditeurs naïfs, et si tous les auditeurs ont des compétences similaires pour accéder à ces indices. De nombreuses études (par ex. [10]) ont montré que le décodage des expressions émotionnelles devait être considéré comme un processus multimodal. De plus Aubergé et al. [3] ont montré que l'information acoustique est intégrée dans le décodage visuel de l'information affective, quand bien même les expressions faciales sont porteuses d'informations fortes. Nous considérons donc des expressions mulimodales, en cherchant à évaluer la part d'information portée par chaque modalité.

2. PAROLE EXPRESSIVE ACTÉE ET SPONTANÉE

Les énoncés utilisés ont été extraits du corpus expressif Sound Teacher/E-Wiz [1]. Ce corpus a été enregistré avec une technique de Magicien d'Oz, dans laquelle le sujet croit interagir avec une interface personne-machine complexe alors que le comportement de l'application est en réalité contrôlé à distance par l'expérimentateur. Il s'agit ici d'une application présentée comme un logiciel novateur d'aide à l'apprentissage des langues étrangères basé sur un système de reconnaissance vocale, pour lequel de derniers tests seraient nécessaires avant sa commercialisation. L'interaction avec le système est contrainte par un langage de commande composé des mots

monosyllabiques français [bʁik], [ʒon], [ʁuʒ], [sabl] et [vɛʁ] ainsi que de la commande [paʒsɥivãt], afin de pouvoir collecter des énoncés identiques exprimant diverses valeurs affectives. Les performances attribuées aux 17 sujets ayant participé à l'expérience ont été manipulées selon un scénario prédéfini afin d'induire chez eux des émotions positives puis négatives. Les productions recueillies ont été étiquetées dans un premier temps par les sujets eux-mêmes à partir de l'enregistrement vidéo avant une étape de validation perceptive. Un protocole spécifique a été mis en place pour les 7 sujets qui étaient également acteurs, à qui il a été demandé immédiatement après l'enregistrement de reproduire les affects ressentis lors de l'expérience sur les mêmes énoncés, les expérimentateurs mettant l'accent sur le fait que ces affects devaient être exprimés de la façon la plus similaire possible au ressenti dans l'expérience. Les acteurs recrutés pour cette tâche pratiquent le théâtre de rue et/ou d'improvisation, et ont déclaré que le dispositif expérimental leur fournissait des conditions optimales pour développer leur jeu d'acteur.

Une première étude utilisant des stimuli actés et spontanés issus de ce corpus a été menée par Laukka et al. [9]. 193 énoncés actés et spontanés produits par 6 acteurs (3 hommes, 3 femmes) et exprimant des émotions liées aux classes de la peur, de la colère et de la joie ont été validés et évalués en termes d'intensité émotionnelle par des auditeurs francophones en condition audio seule dans un prétest, montrant une intensité émotionnelle perçue plus importante plus les énoncés actés que spontanés, qui confirme les résultats de Wilting et al. [12] en condition visuel seul. 24 paires de stimuli constituant un sous-ensemble des production de 4 acteurs (2 hommes notés M1 et M2 et 2 femmes notées F1 et F2) évalués dans ce prétest ont été retenues pour la tâche de discrimination présentée ici.

3. EVALUATION PERCEPTIVE

Les 24 paires de stimuli ont été présentées aux juges séquentiellement, avec une latence de 1,5 secondes, selon 3 conditions de présentation : audio seul (A), visuel seul (V) et audiovisuel (AV). Les stimuli ont été présentés regroupés par condition et triés aléatoirement au sein de chaque condition, la moitié des juges traitant la condition A avant V et vice-versa, tandis que la condition AV était toujours présentée en dernier. De plus chaque paire était présentée 2 fois dans chaque condition, de façon à présenter l'énoncé spontané en 1ère et 2ème position afin de compenser un effet éventuel de l'ordre de présentation. Les juges devaient donc évaluer 144 paires dans l'ensemble de l'expérience. Après avoir été informés du contexte de l'enregistrement du corpus, il leur était demandé pour chaque paire d'indiquer à l'aide d'un curseur (11 points) allant de « certainement le premier » à « certainement le deuxième » lequel des 2 stimuli présentés correspondait à l'expression spontanée. 33 juges francophones (15 hommes, 18 femmes, âgés en moyenne de 33,1 ans) ont participé à cette évaluation perceptive, d'une durée moyenne de 25 minutes.

4. ANALYSE STATISTIQUE

Les positions du curseur ont été converties en scores de discrimination (bonne ou mauvaise réponse) et de confiance, similairement à la méthode utilisée par Bänziger [4]. Les scores moyens de discrimination par condition et par locuteur sont présentés figure 1. Ces scores ne sont que modérément corrélés au taux de confiance quelque soit la condition (A : r=.408 ; V : r=.690, AV : r=.583 ; global : r=.622). Ces valeurs ont été analysées au moyen d'une ANOVA de mesures répétées avec le locuteur, le juge, la classe émotionnelle, la condition de présentation, la longueur de l'énoncé et l'ordre de présentation comme facteurs fixes. Malgré la faible corrélation entre discrimination et confiance, la plupart des effets statistiques observés sur les scores de discrimination l'ont aussi été sur le taux de confiance, les résultats présentés ici sont donc centrés sur les scores de discrimination.

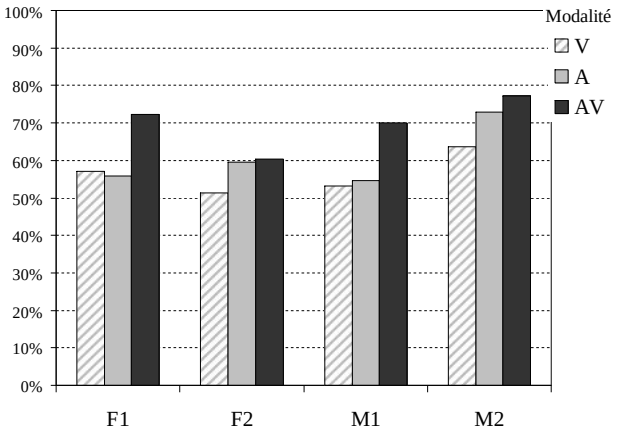

Figure 1 : Discrimination moyenne par locuteur et par condition de présentation

4.1. Un fort effet inter-juge

Le résultat le plus remarquable est l'effet inter-juge ($p<.001$), qui confirme les résultats obtenus sur l'amusement par Aubergé et al. [3] : les performances des auditeurs s'étagent en effet de 32,7% à 80,6% de paires correctement discriminées. Malgré cette grande variabilité dans les performances des juges, ces derniers n'ont pas montré de préférence marquée pour l'un ou l'autre locuteur indépendamment de la qualité de leurs performances d'acteurs, comme le montre la valeur élevée du coefficient alpha de Cronbach (0,867) qui indique que les compétences des juges ont été consistantes d'un locuteur à l'autre.

La figure 2 présente la distribution des scores de discrimination obtenus par les différents juges. 70% des auditeurs ont été capables de discriminer correctement plus de la moitié des paires présentées en condition V, 79% en condition A et 85% en condition AV.

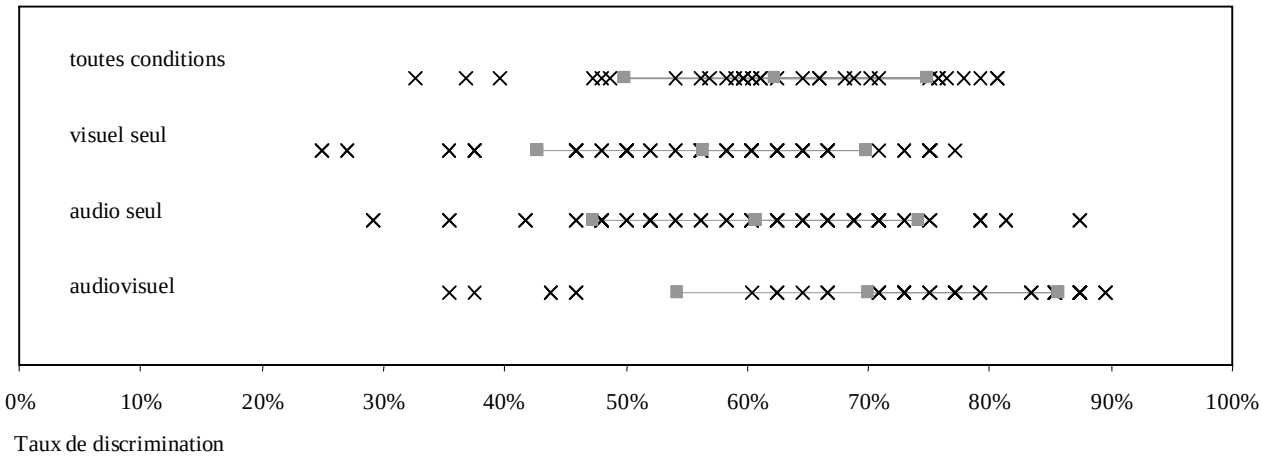

Figure 2 : Distribution des scores de discrimination individuels par condition de présentation. Les lignes et carrés gris indiquent la moyenne et l'écart type pour chaque condition.

4.2. Effet de la modalité et du locuteur

Un effet significatif de la condition de présentation ($p<.001$) a été observé, avec un gain significatif de discrimination pour la condition AV par rapport à A et V ($p<.001$ dans les 2 cas), tandis que l'avantage de A par rapport à V était non-significatif. Cependant cet avantage global pour la condition AV n'est pas consistant pour tous les locuteurs, comme illustré par la figure 1 : aucun effet de la condition n'a été observé pour la locutrice F2, de même que le gain entre A et AV s'est avéré non significatif pour le locuteur M2.

Un effet significatif du locuteur ($p<.001$ en conditions A et AV, $p<.05$ pour V) a également été observé. Seules les productions du locuteur M2 ont été mieux discriminées que celles des autres locuteurs ($p<.001$ dans les 3 cas), indiquant des compétences moindres de cet acteur pour simuler des expressions émotionnelles de manière similaire à ses productions spontanées. Bien qu'un nombre important de juges aient considéré la tâche de discrimination plus difficile pour les locuteurs F2 et M1, ce qui s'est traduit par des scores de confiance significativement plus faibles ($p<.001$), les productions des locuteurs M1, F1 et F2 ont été discriminées avec des scores comparables.

4.3. Autres effets

Aucun effet de la classe émotionnelle n'a été trouvé, indiquant des compétences similaires des juges pour discriminer les productions actées vs. spontanées quelle que soit l'émotion exprimée. L'effet de la longueur de l'énoncé est faiblement significatif ($p<.05$) avec une discrimination légèrement meilleure sur les énoncés [paʒsɥivãt] plus longs.

L'effet de l'ordre de présentation des stimuli dans la paire (stimulus spontané en 1ère ou 2ème position) s'est révélé significatif ($p<.001$) en condition V uniquement et uniquement pour les locuteurs M2 (le mieux discriminé) et F2 (la moins bien discriminée), qui étaient également les locuteurs pour lesquels le gain entre les conditions A et AV était non significatif. Bien que les expressions faciales et la gestualité des locuteurs n'aient pas fait l'objet d'une analyse objective et systématique, l'amplitude des gestes de M2 et F2 est de toute évidence plus importante que pour les autres locuteurs. De plus tous deux bougeaient quasi-systématiquement la tête vers le bas pendant leurs productions spontanées. Une possible explication de cet effet de l'ordre de présentation pourrait être la force informative du stimulus visuel présenté en premier.

Bien que les scores de discrimination et de confiance attribués par les juges soient plus élevés pour les femmes, tout particulièrement en condition A, ces différences inter-genre ne sont pas significatives. Quoique la significativité statistique ne puisse être calculée dans ce cas précis, un résultat particulier mérite d'être relevé : pour 2 juges hommes ayant des performances en condition A meilleures que la moyenne mais parmi les moins bonnes en condition V, l'interprétation erronée de l'information visuelle semble avoir largement grevé leurs performances en discrimination en condition AV. En effet pour ces 2 juges les scores de discrimination en condition AV sont de plus de 20% inférieurs aux scores en condition A.

Les corrélations entre les différences de durée dans chaque paire (de -480 ms à 760 ms) d'une part, et les scores moyens par paire de discrimination et de confiance d'autre part ont été calculées afin d'étudier la possible influence de ces différences de durées sur les performances en discrimination. Cependant les très faibles valeurs de ces corrélations (respectivement $r=.047$ et $r=.037$) suggèrent que cet indice n'a pas été utilisé par les juges.

4.4. Rôle de l'intensité émotionnelle perçue

Afin d'évaluer le rôle de l'intensité émotionnelle perçue dans la discrimination des expressions actées vs. spontanées, les corrélations partielles pour chaque condition entre la différence d'intensité perçue dans la paire extraite de [9] et les scores de discrimination (respectivement de confiance) ont été calculées. Ces corrélations partielles sont présentées dans la table 1. Ces intensités perçues ayant été attribuées en condition audio seul, il n'est pas surprenant d'observer des corrélations plus élevées dans cette condition.

Table 1 : Corrélations entre différence d'intensité émotionnelle perçue et discrimination (resp. confiance)

condition	A	V	AV	global
discrimination	r=.745	r=.131	r=.335	r=.415
confiance	r=.402	r=.147	r=.283	r=.250

Bien que les paires pour lesquelles la différence d'intensité perçue en condition audio-seul est la plus importante soient également les mieux discriminées, indiquant que cette caractéristique pourrait être un indice fort utilisé prioritairement lorsqu'il est possible d'y accéder, les juges ne semblent pas s'être appuyés uniquement sur cet indice. En effet parmi les 3 paires pour lesquelles la différence d'intensité perçue était la plus faible, une seule a été discriminée par moins de la moitié des juges en condition A, les deux autres étant discriminées par plus de 60% des juges. À l'inverse, deux paires pour lesquelles la différence d'intensité perçue était plus importante ont été parmi les moins fréquemment discriminées.

5. CONCLUSION

Les résultats obtenus suggèrent une capacité globale des auditeurs naïfs à discriminer des expressions émotionnelles multimodales actées vs. spontanées, sans effet de l'émotion exprimée mais avec un fort effet inter-juge qui pourrait être lié à la notion d'intelligence affective. Au-delà de la simple discrimination, la question d'une éventuelle variabilité des compétences individuelles pour identifier le recours au processus de simulation (qui est dans nos hypothèses [2] à la base de l'expression des affects sociaux ou attitudes) reste ouverte.

Les différences d'intensité émotionnelle perçue, mesurée auparavant en condition audio [9] sur les stimuli individuels, semblent en mesure d'expliquer une part importante de la facilité à discriminer sans pour autant expliquer toute la variabilité. Afin de permettre une mise en correspondance plus systématique entre différence d'intensité émotionnelle perçue et discrimination des expressions actées vs. spontanées, une évaluation perceptive de l'intensité émotionnelle reprenant les mêmes stimuli et l'ensemble des conditions de présentation est actuellement en préparation. De plus, une analyse acoustique et visuelle des stimuli utilisés est en cours, basées des hypothèses fortes quant à l'ancrage temporel des expressions volontaires vs. involontaires [2].

Bien que les acteurs choisis ne soient certainement pas parmi ceux reconnus comme les meilleurs, 3 sur 4 ont été en mesure de piéger les juges les moins performants. Sans préjuger de la capacité d'autres acteurs à piéger un nombre plus important de juges, la parole actée, fréquemment utilisée comme référence pour la modélisation de phénomènes relatifs à l'expression d'émotions involontaires, pourrait être reconsidérée en tenant compte de cette capacité de discrimination. Au-delà de la discrimination la question de la variabilité des performances en identification de la simulation (qui est dans notre hypothèse à la base des expressions d'affects sociaux) reste ouverte, ainsi que celle de son possible lien avec le quotient affectif.

BIBLIOGRAPHIE

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

[2] V. Aubergé. A Gestalt morphology of prosody directed by functions: the example of a step by step model developed at ICP. *Proc. 1st Speech Prosody*, pages 151-155, 2002.

[3] V. Aubergé and M. Cathiard. Can we hear the prosody of smile? *Speech Communication* 40 (2) : 87-97, 2003.

[4] T. Bänziger. *Communication vocale des émotions. Perception de l'expression vocale et attributions émotionnelles.* Thèse de doctorat, Université de Genève, 2004.

[5] N. Campbell. Databases of Emotional Speech. *Proc. ISCA WS on Speech and Emotions*, pages 34-38, 2000.

[6] F. Enos and J. Hirschberg. A Framework for Eliciting Emotional Speech: Capitalizing on the Actor's Process. *Proc. WS Corpora for Research on Emotion and Affect*, pages 6-10, 2006.

[7] A.R. Damasio. *Looking for Spinoza. Joy, Sorrow an the Feeling Brain.* Orlando:FL/Harcourt, 2003

[8] I. Fonagy. *La vive voix.* Paris:Payot, 1983.

[9] P. Laukka, N. Audibert and V. Aubergé. Graded structure in vocal expression of emotion: What is meant by "prototypical expressions"? *Proc. WS on Paralinguistic Speech*, pages 1-4, 2007.

[10] K. R. Scherer and H. Ellgring. Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns? *Emotion*, 7(1) : 158-171, 2007.

[11] K. R. Scherer. Appraisal considered as a process of multi-level sequential checking. In *Appraisal processes in emotion: Theory, Methods, Research*, Scherer K. R., Schorr A., & Johnstone T. (eds.), Oxford Univ. Press, pages 92-120, 2001

[12] Wilting, J.; Krahmer, E.; Swerts, M., 2006. Real vs. acted emotional speech. *9th INTERSPEECH*, Pittsburgh, PA, USA (CD-ROM proceedings).