De E-Wiz à C-Clone

Recueil, modélisation et synthèse d'expressions authentiques

Véronique Aubergé — Nicolas Audibert — Albert Rilliard

Institut de la Communication Parlée
UMR CNRS 5009
INPG-Université Stendhal
1180, avenue Centrale – BP 25
F-38040 Grenoble cedex 9

{Veronique.Auberge, Nicolas.Audibert, Albert.Rilliard}@icp.inpg.fr

RÉSUMÉ. *Différents niveaux d'affects sont exprimés dans différents niveaux du traitement de la parole : les expressions des émotions, relevant d'un contrôle déclenché involontairement, les expressions des attitudes et des intentions du locuteur et les stratégies expressives métalinguistiques. C-Clone (Communicative Clone) est présenté comme une architecture cognitive interactive de la communication expressive. Des corpus expressifs authentiques et néanmoins contrôlés ont été recueillis grâce à une plate-forme expérimentale de type magicien d'Oz (E-Wiz). Le signal acoustique, la vidéo et certaines mesures physiologiques ont été enregistrés synchroniquement. Un sous-ensemble auto-annoté d'expressions émotionnelles a permis de valider la modélisation de la prosodie affective en contours gradients. Des expériences perceptives indiquent en outre qu'aucune dimension acoustique prosodique n'est spécifique à une valeur d'émotion.*

ABSTRACT. *Different levels of affects are expressed in different levels of speech processing: expressions of emotions linked to an involuntarily triggered control, expressions of the speaker's attitudes and intentions and meta-linguistic expressive strategies. C-Clone is presented as an interactive cognitive architecture of the expressive communication. Authentic but controlled expressive corpora have been collected using a Wizard of Oz experimental platform (E-Wiz) The acoustic as well as video and some physiological signals were recorded synchronously. An auto-annotated subset of emotional expressions made possible the validation of the modeling of affective prosody into gradient contours. Moreover perceptive experiments show that no prosodic dimension is specific of an emotional value.*

MOTS-CLÉS : *affects, émotions, attitudes, intentions, expressivité, agent communicant expressif, prosodie affective, parole expressive, expressions, expérimentation, corpus authentiques, contours gradients.*

KEYWORDS: *affects, emotions, attitudes, intentions, expressivity communicant expressive agent, affective prosody, expressive speech, expressions, experimentation, authentic corpora, gradient contours.*

RSTI - RIA – 20/2006. Interaction émotionnelle, pages 499 à 527

1. Introduction

La parole est une modalité incontournable de l'expression des affects de l'humain. Différents processus de la parole véhiculent les informations relatives à différents niveaux d'affects (les émotions, mais aussi les attitudes et l'expressivité). En cohérence, et par extension à plusieurs hypothèses séparatrices des traitements affectifs véhiculés par la parole, en particulier (Fonagy, 1983) et (Léon, 1993), l'hypothèse principale qui sous-tend l'ensemble des travaux présentés ici est que le matériel acoustique de la parole en tant que vecteur de la communication face-à-face d'un sujet avec un ou des agents (humains ou virtuels), intègre différents niveaux cognitifs d'informations affectives.

1) Les émotions sont déclenchées par un contrôle involontaire et sont exprimées dans et par la voix (la granularité de la voix n'étant pas considérée comme celle de la parole, vecteur du langage) ; même lorsque la face véhicule une information plafond dans la modalité visuelle, largement transmise dans l'acoustique de la parole à travers la déformation audible de la face (sourire, dégoût), des informations sont spécifiquement véhiculées dans le canal vocal (Aubergé et Cathiard, 2004). Les processus d'expressions des émotions, vraisemblablement innés, sont soumis à l'apprentissage à travers le contrôle de l'intensité de l'expression (de inhibition à exagération). Le temps cognitif dans lequel sont régies les expressions des émotions est organisé par les événements de causalité des variations émotionnelles induisant l'expression, que ces événements soient internes au contexte communicationnel, ou que ces événements soient externes. Ces affects émotionnels sont les plus largement observés et étudiés dans les différentes théories sur les émotions.

2) Les affects intentionnels, les attitudes, sont déclenchés par un contrôle volontaire et sont exprimés directement en particulier par la prosodie, organisée dans le temps cognitif du langage et de la phonologie. Les concepts et les morphologies prosodiques en sont construites par la langue et la culture et sont acquises (installées entre 7 et 11 ans selon Clément (1999)). Ils relèvent d'un contrôle cognitif volontairement décidé. La parole véhicule ainsi, en plus du « quoi dire » de l'énoncé produit par le locuteur, ses attitudes et ses points de vue (doute, autorité, confiance, politesse...). Lorsque le locuteur décide de ne pas donner son point de vue et attitude (la valeur de l'énonciation est alors réduite à la modalisation : par exemple dans des contextes sociaux ou le locuteur est diffuseur d'informations – journaux, radio, TV), il s'agit d'une attitude particulière qui est précisément de ne pas informer de son attitude. Ce type d'affects est celui qui occupe le plus largement le canal de la parole, et constitue la partie essentielle, en occurrence et en variance, de la communication parlée (Campbell, 2004 ; Aubergé *et al.*, 1997 ; Aubergé, 2002 ; Morlec *et al.*, 2001 ; Shochi *et al.*, 2005 ; Wichmann, 2002). Cependant, lorsqu'un affect émotionnel est exprimé, même ou précisément s'il est rare, sa présence est fondamentale et il ne doit en aucun cas être négligé dans une perspective de traitement automatique, même en dehors des contextes à forte induction émotionnelle.

3) Le niveau le plus complexe des affects est celui de l'expressivité langagière : il résulte de métaprocessus sur l'ensemble des structures linguistiques (la prosodie, le lexique, la morphologie et la syntaxe). Comme la morphologie et la syntaxe, la prosodie remplit des fonctions linguistiques comme la segmentation et la hiérarchisation de l'énonciation, la modalisation, la focalisation (Rossi, 1999 ; Aubergé, 2002). La prosodie est donc bien ou mal formée linguistiquement, et il existe plusieurs bonnes formes possibles cohérentes avec une même syntaxe d'énonciation (Rilliard et Aubergé, 2003). Le traitement métalinguistique expressif prosodique est le choix d'une stratégie parmi l'ensemble de ces « paraphrases prosodiques ». Ces métastructures véhiculent des informations affectives, souvent désignées à tort comme des éléments flous de la stylistique.

Nous présentons en première partie l'architecture cognitive C-Clone (Communicative Clone), que nous proposons comme un modèle d'intégration des différentes compétences linguistiques et affectives d'un agent communicant affectif anthropomorphique. C-Clone est régi dans le temps événementiel (fonction émotionnelle) et dans le temps linguistique (fonction attitudinale/intentionnelle et fonction expressive). C-Clone pose le problème de la « corporéisation » comme un problème de « contrôles-contraintes » et non comme une motricité du corps commandé en sortie d'un processus logique. Les expressions affectives sont produites, pour ce qui concerne la parole, dans l'hypothèse théorique de la boucle de la perception-action. Ainsi, une question centrale est de comprendre et modéliser comment les différents traitements affectifs sont implémentés dans le même matériel acoustique. La récupération de la prosodie émotionnelle *vs.* la prosodie affective intentionnelle est-elle basée sur une séparation des paramètres acoustiques ? Certains paramètres sont-ils réservés au traitement de la prosodie émotionnelle *vs.* affective intentionnelle/expressive ? (Bänziger *et al.*, 2003) Cette récupération est-elle résolue par une différence dans la nature des traitements morphologiques : le traitement gradient serait réservé aux expressions émotionnelles (Scherer *et al.*, 1984), tandis que la modélisation prosodique de type linguistique serait réservée aux expressions des attitudes car régies dans le temps linguistique.

Ces hypothèses doivent être testées dans un paradigme expérimental qui permet d'accéder à des données représentatives de chaque niveau d'affects. Ce paradigme s'inscrit dans une boucle méthodologique classique : recueil de données représentatives des hypothèses à évaluer, analyse, modélisation, simulation, évaluation, et retour éventuel sur les hypothèses (cf. figure 1). La seconde partie de cet article décrit une plateforme et un protocole à partir desquels ont pu être recueillies des expressions authentiques et néanmoins contrôlées en condition de laboratoire. La collecte de données émotionnelles pose en soi des questions méthodologiques qui sont d'abord discutées. La méthodologie d'auto-annotation mise en œuvre dans cet article fait intervenir la compétence « naïve » de mémoire et de perception du sujet sur ses propres productions. La plateforme E-Wiz *(Expressive-Wizard of Oz)* a été développée pour mettre au point des scripts destinés à l'enregistrement d'expressions authentiques mais contrôlées d'expressions

émotionnelles, attitudinales et expressives. Ensuite le scénario *Sound Teacher* est détaillé, ainsi que les corpus enregistrés à l'aide de ce scénario, en particulier les corpus de parole émotionnelle enregistrés pour plusieurs locuteurs dans les mêmes situations authentiques, avec des traitements émotionnels comparables.

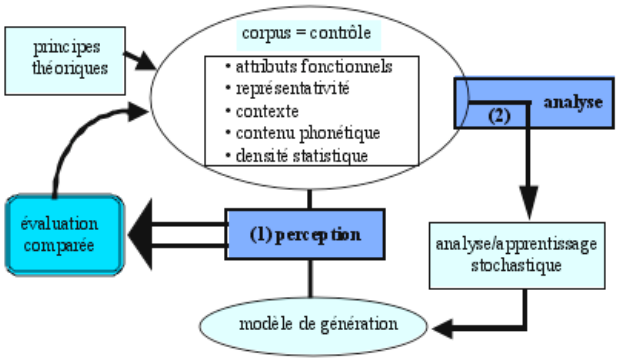


Figure 1. *Méthodologie expérimentale*

La dernière partie présente les analyses acoustique et perceptives menées sur le sous-corpus des énoncés représentant le plus isolément les expressions émotionnelles (attitudes supposées gelées et pas de structure morpho-syntaxique). Une comparaison est présentée entre une approche de modélisation gradiente *vs.* par contours. Les expériences perceptives valident d'abord la pertinence du corpus audiovisuel puis fournissent des échelles de référence pour l'évaluation des poids relatifs des différents paramètres acoustiques véhiculant les expressions.

2. C-Clone : une architecture coopérative pour la communication expressive

Le système de communication d'un ACA (Agent Communicant Animé) est conçu dans C-Clone (figure 2) selon un postulat anthropomorphique de simulation des compétences communicatives de l'humain. Ce système, qui est dirigé par les buts de l'interaction verbale, utilise un ensemble de fonctions globalement régies par le système. Selon notre hypothèse cognitive, ce système est une architecture modulaire, non pas dans une structure fodorienne (pas de module central dans C-Clone), non pas dans une architecture modulaire en graphe de type Levelt, mais dans une organisation coopérative, typiquement multi-agent, et dans laquelle les fonctions émergent des interactions entre les modules. Chaque module (lexique, morphologique, syntaxe, prosodie) peut ainsi se définir par une morphologie autonome (mais non indépendante), des degrés de liberté et des contraintes qui lui sont spécifiques. La cohérence des différents modules qui coopèrent à l'encodage d'un même domaine

fonctionnel est assurée par une stratégie globale d'instanciation de chaque fonction, dont les traces saillantes sont les rendez-vous structurels (Aubergé, 1992). Ainsi pour une fonction communicative donnée, linguistique ou affective, plusieurs modules coopèrent (stratégie intrafonction), chacun étant un système autonome contrôles/contraintes. Ainsi la fonction de segmentation/hiérarchisation de l'énonciation (couper en syntagmes et les ordonner paradigmatiquement) est prise en charge universellement par la syntaxe bien sûr, mais aussi par la prosodie (Aubergé, 1992 ; Rilliard et Aubergé, 2003), et par la gestualité (McNeil, 1992). La fonction de modalisation est réalisée universellement par la syntaxe, la prosodie et peut-être la gestualité. La fonction de focalisation est réalisée également par la syntaxe pour certaines langues (processus d'extraction) et par la prosodie, l'intensification s'ajoutant à la focalisation par la morphologie (adverbes intensifieurs, choix lexicaux) et la prosodie (Aubergé et Rilliard, 2005). Il en est de même pour les attitudes et l'expressivité. Pour l'ensemble de ces contrôles volontaires, le choix de véhiculer une information fonctionnelle dans un ou plusieurs de ces modules, avec des rendez-vous plus ou moins réguliers, est en soi une information véhiculée méta-linguistiquement par la stratégie intrafonctionnelle. Ainsi lorsque l'on modélise la prosodie d'un ACA anthropomorphique comme C-Clone, sa compétence prosodique ne s'arrête pas à la granularité du module prosodique, mais réside également dans la coopération entretenue par le module prosodique avec l'ensemble des autres modules pour toutes les fonctions. 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, plus d'intelligibilité de focalisation) est aussi un vecteur d'information fondamental (stratégie interfonction).

Quant à la fonction émotionnelle, traitée dans le temps primaire des événements émotionnels, que le temps linguistique de la communication encapsule, elle contrôle, involontairement, par des processus automatiques, une répartition de l'information dans trois modalités : la face, le geste, la voix. La question que nous nous posons est de déterminer quel degré de contrôle possède le système communicatif sur ces expressions. Pour une variation donnée des états émotionnels de l'agent, nous voulons tester trois types de contrôles selon leur validité anthropomorphique :

1) le seul choix de l'agent est un potentiel d'inhibition de l'expression également répartie sur chacune des trois modalités d'expression (redondance totale : mêmes nature et valeur d'émotion sur les trois modalités) ;

2) un potentiel inhibiteur est donné à chaque modalité (redondance modulée), ainsi qu'une même valeur émotionnelle guide chaque modalité mais avec une gradience spécifique, ce contrôle étant alors une information (e.g. plus de joie dans la voix que le visage aurait un sens affectif) ;

3) enfin on doit se demander s'il est possible d'envoyer des informations complémentaires (*i.e.* non redondantes) dans chaque modalité plutôt que la redondance totale du mélange souvent complexe d'émotions exprimées.

Ces questions font partie des motivations des travaux expérimentaux que nous présentons ci-dessous, et justifient la mesure précise des expressions faciales. Même

si notre but est avant tout de modéliser la parole et la voix affectives, leur relation aux autres modules et modalités est incontournable puisque porteuse d'information. Ainsi lorsque nous analysons la voix, nous la considérons comme intégratrice de plusieurs vecteurs de contrôle : le contrôle spécifique de la voix, le contrôle de la face rendu audible par une intégration visio-audible (Aubergé et Cathiard, 2004), et le contrôle physiologique qui peut modifier globalement tous les articulateurs (e.g. les indices somatiques liés à la peur) et être indirectement audible.

Les choix stratégiques, comme l'ensemble des décisions linguistiques (y compris les affects du temps linguistique, *i.e.* les attitudes et l'expressivité) sont des prises de décisions rationnelles. Nous pouvons donc prendre l'hypothèse de Damasio au compte des décisions interactionnelles qui proposent les états émotionnels comme évaluateur de ces décisions (théorie de l'évaluation ou « appraisal » de Lazarus ou Scherer) ou préparateur à l'action (Frijda, 1987).

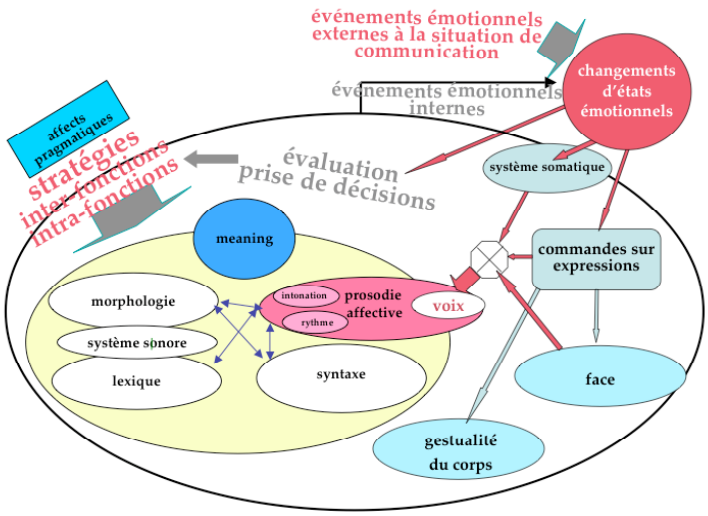


Figure 2. *Architecture cognitive de C-Clone*

Dans C-Clone, et c'est là une hypothèse fondamentale, les modules ne sont pas réductibles à des traitements de logique calculable qui activeraient en sortie pour la parole, la face et le corps des commandes de motricité ou de pseudo-motricité. Le modèle de prosodie que nous proposons intègre la morphologie physique des expressions affectives dans sa description fonctionnelle. La multimodalité des expressions est posée comme un problème de cohérence et non comme un problème de corrélation, et s'inscrit dans la cadre de la théorie de la perception-action, dans laquelle agir et percevoir partagent la même représentation pour un geste biologique

sémantique. C-Clone pose le problème de la « corporéisation » comme l'intégration du geste dans la représentation interne du modèle.

3. E-Wiz : capture *in vitro* d'expressions émotionnelles authentiques

Les corpus de parole émotionnelle peuvent être classés selon trois axes orthogonaux : 1) les méthodes *in vivo* vs. *in vitro* – ces dernières faisant référence aux enregistrements de laboratoire, qui impliquent un enregistrement de parole dans des conditions optimales (Campbell, 2000) ; 2) le *degré de contrôle* du non-contrôle au contrôle total de certaines caractéristiques de la parole (la situation d'interaction, le contenu linguistique ou phonétique…) ; 3) les protocoles *actés* vs. *authentiques* (avec des niveaux intermédiaires par l'induction et l'élicitation). Ces dimensions peuvent être superposées : par exemple, des données authentiques sont plus facilement recueillies *in vivo* et sans contrôle par l'observateur. Un état de l'art sur le recueil de corpus émotionnels est donné par Douglas-Cowie *et al*., (2003).

Dans le cadre d'enregistrement *in vivo*, un contrôle a très rapidement été introduit par la présence d'un compère (Williams et Stevens, 1972 ; Scherer *et al.*, 1984). Dans le cas des corpus *in vitro*, les acteurs ont été largement utilisés, avec des techniques plus ou moins élaborées d'éllicitation (Amir *et al.*, 1998 ; Mozziconacci, 1998 ; Scherer, 2001). D'autres principes d'induction se basent sur la présentation de stimuli, par exemple d'images à fort contenu émotionnel, juste avant la production de parole par un locuteur non professionnel, afin d'étudier les changements somatiques induis chez les sujets, et les expressions ne sont alors pas mesurées. Par ailleurs, Iida (2002) a demandé à des locuteurs non professionnels de lire des textes fortement chargés sur le plan émotionnel.

Enfin d'autres études, moins nombreuses, ont été destinées au recueil d'expressions authentiques tout en conservant le degré de contrôle élevé propre aux enregistrements *in vitro*, grâce à la présence d'un compère et à l'utilisation de tâches précises qui permettent de contraindre la nature des énoncés recueillis et d'induire les variations émotionnelles attendues. Ainsi Johnstone *et al*. (1999) ont utilisé un scénario basé sur un jeu vidéo, et Kaiser *et al.* (1994) ont imaginé un scénario d'interaction basé sur une tâche informatique.

Les corpus non actés posent le problème fondamental de l'annotation : concernant les affects volontaires, « conventionnalisés » par la langue et la culture, il est légitime de supposer qu'un annotateur humain puisse se placer en situation méta-communicative d'observation, et prétende à la compétence d'expert, c'est-à-dire sorte de ses compétences naturelles de communicant pour développer une compétence objectivée. Il est par contre discutable qu'un humain en situation d'observation d'expression d'émotion (contrôle involontaire, et dont la perception met en œuvre des processus complexes comme l'empathie) puisse être considéré comme expert, puisqu'il utilise directement, pour l'annotation, ses compétences « naïves » d'humain dans une situation écologique de communication qui définit

comme l'observation « émotionnelle », par un humain, d'autres humains en situation émotionnelle. De plus, dans les corpus recueillis sur le vif (comme des entretiens journalistiques de personnes en situations extrêmes), 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 annotateur, peu de contraintes lui permettant alors d'éviter les « erreurs » naturelles ou la subjectivité qui permet à un humain par exemple d'établir des prédictions. C'est pourquoi la plupart des travaux d'annotation établissent soit des tests de cohérence entre plusieurs experts, soit vérifient l'annotation par des mesures perceptives effectuées par un panel générique d'auditeurs naïfs (Martin *et al.*, 2005). Mais ces contrôles sont difficiles à mener systématiquement sur de grands corpus.

3.1. *Parole actée vs. authentique*

La parole actée est spécifique par divers aspects. Les acteurs expriment des émotions dont la naturalité est variable selon le contexte artistique (théâtre, cinéma, « rue » – improvisation), les méthodes et les cultures, et qui peuvent être fort éloignées d'une imitation ou simulation d'expression authentique. Le fait que ces productions soient aisément identifiables dans des tests perceptifs, ne signifie pas qu'elles soient identiques à une production non actée. De tels effets peuvent être attendus (avec même de meilleurs scores qu'en parole non actée) pour une parole actée très caricaturale et stéréotypée. Si les énoncés sont produits avec une méthode de simulation (prétendre est une compétence acquise par l'enfant) sans ressenti mémoriel, ce processus renvoie à une classe de processus que nous ne classons pas dans les émotions, mais dans les processus intentionnels. Par contre, si l'on sélectionnait uniquement des acteurs dont la méthode est de « re-ressentir » une émotion vécue, la compétence intrinsèque de l'acteur, et la qualité de ses performances en laboratoire, loin du contexte de mise en scène favorisant son jeu, ne permet pas de maîtriser, ni d'évaluer, la qualité d'imitation des ses productions. Ainsi, une première expérience menée par Aubergé et Cathiard (2003) montre que l'amusement acté peut être discriminé de l'amusement non acté. Cette étude a surtout mis en évidence un effet interjuge : certains juges n'identifiaient pas les acteurs, alors que d'autres les identifiaient avec un bon score.

3.2. *Le paradigme du Magicien d'Oz pour le contrôle des données*

Si l'on considère les trois niveaux imbriqués d'expression des affects, le niveau le plus bas, celui des émotions, peut être isolé *(ceteris paribus)* des niveaux supérieurs d'affects, en gelant la variabilité des attitudes et intentions et en contraignant les énoncés à des unités lexicales choisies pour constituer à elles seules les énoncés. Il est très difficile de « rencontrer » de telles productions dans des situations non conçues dans ce but. La méthode la plus couramment mise en œuvre

est la production par des acteurs, mais nous avons vu plus haut pourquoi nous l'avons rejetée, pour nous tourner vers la technique du magicien d'Oz.

Le paradigme du magicien d'Oz, largement utilisé dans le cadre de l'évaluation d'interfaces multimodales, consiste en l'imitation par un complice humain, le « magicien », du comportement d'une interface personne/machine complexe. Le sujet croit donc qu'il communique avec un ordinateur. Pour induire des comportements émotionnels, il faut soit choisir une tâche qui provoque des réactions émotionnelles dans son déroulement normal, soit perturber le déroulement normal d'une tâche afin que ces perturbations induisent des expressions émotionnelles.

La solution que nous avons retenue pour geler les attitudes et l'expressivité des locuteurs est de les mettre en communication avec la machine, en leur donnant un langage de commande réduit, en ne donnant jamais la parole à la machine, et en ne donnant pas à l'humain la possibilité d'influer sur le déroulement du scénario par ses attitudes ou par la communication de ses intentions. Des scénarios qui au contraire sont destinés à recueillir des attitudes mettront par exemple en scène le sujet communicant avec un humain, avec comme but commun la résolution d'une tâche donnée sur la machine. Là encore le langage de commandes utilisé ne doit pas permettre l'usage de l'expressivité linguistique. Enfin, pour collecter des corpus d'expressivité, on laisse au sujet, pour une tâche définie, toute latitude dans l'utilisation des outils linguistiques.

Quel que soit le but, les données seront ainsi authentiques, mais leur contenu sera contrôlé et plusieurs locuteurs pourront être enregistrés dans des conditions *in vitro* identiques, provoquant les mêmes réactions. Il est même possible de reproduire l'expérience sur diverses langues et cultures. Cette méthode est donc adaptée à des études à visées aussi bien fondamentales qu'applicatives.

Le préambule nécessaire à la mise en œuvre de tels scénarios d'induction est de choisir la tâche et les sujets afin que leur implication soit directement la tâche elle-même (et non des motivations indirectes comme une rétribution, une note, la curiosité, etc.).

3.3. *E-Wiz: une plateforme dédiée*

Afin de mettre en œuvre des expériences basées sur le paradigme du Magicien d'Oz et destinées à collecter des corpus de parole émotionnelle authentique, une plateforme spécifique, E-Wiz (figure 3), a été développée à l'ICP. Cette plateforme, développée en langage Java avec une architecture client/serveur (Rebreyend, 2002), permet à un utilisateur ne possédant pas de compétences informatiques particulières d'élaborer graphiquement de nouveaux scénarios d'induction. La plateforme se décompose en trois applications distinctes, dont un éditeur dédié à la mise au point des scénarios. L'éditeur permet de générer automatiquement des scripts de configuration décrivant entièrement le comportement des applications client et

serveur pour un scénario donné. Ces applications utilisent directement les scripts générés par l'éditeur pour la phase d'enregistrement effectif des corpus.

Les scénarios élaborés à l'aide de ce logiciel sont constitués d'un ensemble de pages reposant sur divers types de données multimédia, tels que texte, images et sons. Les images ainsi que les zones d'affichage de texte peuvent être déplacées par le magicien côté serveur pour faire apparaître au sujet une sorte de diaporama côté client. Afin de faciliter la disposition des objets avec l'éditeur, ce dernier a été doté d'une interface conviviale. Ainsi, des fonctionnalités d'édition et de traitement de texte ont été implémentées, permettant une utilisation intuitive de l'application. De plus, la tâche du magicien peut être allégée en automatisant le comportement de certains objets. Par exemple, la lecture des sons peut être liée à l'ouverture de certaines pages, et les déplacements des objets peuvent être recalculés côté client afin de paraître provoqués par la machine. De plus, des compteurs automatiques, dont le comportement peut être prédéfini, peuvent également être intégrés aux pages.

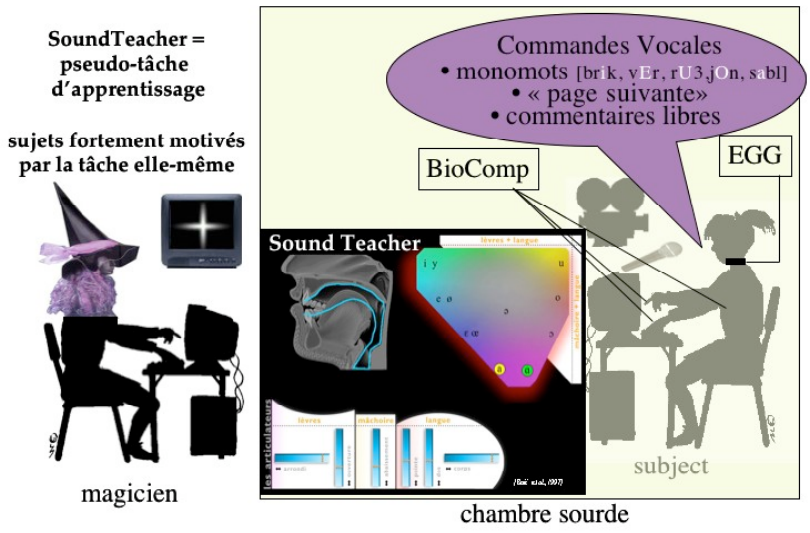


Figure 3. *La plateforme expérimentale E-Wiz*

La plateforme E-Wiz est disponible gratuitement pour un usage non-commercial sur demande à l'un des auteurs.

3.4. *Le scénario Sound Teacher'*

Les scénarios d'E-Wiz développés pour le recueil de parole émotionnelle se basent sur le même principe : le sujet doit interagir avec l'ordinateur à l'aide d'un

langage de commandes. L'utilisation d'un lexique restreint permet d'obtenir différents états émotionnels exprimés sur les mêmes énoncés, afin de faciliter les analyses acoustiques ultérieures. Un premier scénario conforme à ces principes et reposant sur des tests de logique de type QI, *Top Logic*, a tout d'abord été développé mais s'est avéré insuffisamment motivant pour les sujets.

Le second scénario, *Sound Teacher*, se présente comme un logiciel permettant au sujet d'améliorer son apprentissage phonétique des langues étrangères. Les sujets sont choisis en raison de leur grande motivation pour cette tâche. La méthode d'apprentissage est supposée reposer sur des découvertes neuropsychologiques relatives à la théorie de la perception/action. Elle se base sur l'apprentissage de quatre paramètres du conduit vocal (ouverture, position avant/arrière, arrondissement des lèvres, centralisation). Les sujets sont entraînés à reconnaître les valeurs des paramètres lors de l'écoute de voyelles, et à les produire. Le scénario est organisé en quatre étapes, du moins difficile au plus accessible du point de vue de la tâche prétexte. La première étape est de vérifier les capacités du sujet pour la production et la perception de voyelles françaises. Un *feedback* artificiellement positif, nettement supérieur à un score fictif moyen, est donné au sujet, qui doit ensuite apprendre des voyelles proches du système vocalique du français. Le score élevé qui lui est attribué (parmi les 5 meilleurs sujets) lui permet alors de passer directement à une phase de généralisation à des sons complexes. Le *feedback* devient alors subitement négatif : le score attribué au sujet est largement inférieur au score moyen des sujets précédents. Le sujet est averti que ces résultats sont anormaux, et que ses compétences pour les voyelles du français doivent être vérifiées à nouveau, car Sound Teacher pourrait les avoir détériorées. La dernière phase est donc similaire à la première, à cette différence près que les stimuli audio ont été modifiés afin de diminuer fortement le contraste perceptif entre les paires de stimuli présentés, forçant le sujet à répondre au hasard, et qu'un score très faible lui est alors attribué. Des commentaires sont en outre demandés régulièrement au sujet. Un scénario particulier a été mis au point pour les sujets acteurs, afin de pouvoir comparer des expressions authentiques *vs*. actées : immédiatement après avoir été piégés par Sound Teacher, l'instruction leur est donnée de reproduire à l'aide de méthodes d'acteurs les émotions ressenties au cours de l'expérience.

Le matériel sonore collecté pour constituer le corpus consiste en la commande [paʒ sɥivãt] (page suivante), et en 5 noms de couleur monosyllabiques (pour éviter les effets de *timing* et de prosodie à long terme) répartis dans l'espace phonologique : [ʁuʒ], [ʒon], [sabl], [vɛʁ], [bʁik], ainsi que les commentaires donnés par le sujet. La comparaison entre le signal électroglottographique (EGG) et l'algorithme de calcul du quotient d'amplitude développé par Mokhtari et Campbell (2003) pour la détection de la *breathy voice* a permis une évaluation de cet algorithme (Rossato *et al.*, 2004).

Parmi les 17 sujets enregistrés jusqu'à maintenant, 7 étaient des acteurs professionnels, pour lesquels un protocole additionnel a été défini : immédiatement après qu'ils aient été piégés à l'aide de Sound Teacher, nous avons demandé à ces

sujets d'exprimer à nouveau à l'aide de méthodes d'acteurs les états émotionnels qu'ils pensaient avoir ressentis au cours de l'expérience. Cette tâche a été effectuée à la fois sur les mots de commande utilisés dans la partie authentique, et sur 10 phrases sémantiquement « neutres », syntaxiquement prototypiques et de longueur variable (3 à 7 syllabes). De plus, les acteurs ont également simulé les expressions des six émotions (joie, colère, peur, tristesse, surprise, dégoût) connues sous le nom de « big six » (Cornelius, 1996).

Les émotions recueillies chez les 17 sujets sont proches de celles attendues (quoique certainement dépendantes du profil psychologique de chaque sujet) : concentration, satisfaction, joie, soulagement, stress, colère, découragement, ennui, angoisse. Un premier étiquetage émotionnel a été réalisé par les sujets eux-mêmes à la suite de l'enregistrement à partir de l'enregistrement vidéo, une partie du corpus ayant ensuite fait l'objet de validations perceptives.

3.5. *Mesures expérimentales*

Les scénarios développés avec E-Wiz ont été mis en œuvre en chambre sourde, afin d'obtenir un enregistrement acoustique de la parole de haute qualité. Certaines mesures de référence sont conservées afin de vérifier la nature, l'intensité et la localisation des variations d'expressions émotionnelles :

– le signal visuel, qui concerne principalement les mouvements de la face et du haut du corps (les sujets sont en effet assis au cours de l'enregistrement) ;

– les signaux bio-physiologiques (rythme cardiaque, réflexe galvanique, amplitude respiratoire, température, Electro-Myo-Graphe) enregistrés à l'aide de l'équipement Pro-Comp ;

– les signaux articulatoires relatifs à la qualité de voix (restreints jusqu'alors au signal Electro-Glotto-Graphique).

Ces signaux peuvent être analysés parallèlement aux analyses perceptives, et constituent les indices principaux du « *timing* émotionnel » pour identifier les instants auxquels les mouvements prosodiques qui qualifient les expressions émotionnelles doivent être mesurés. Le premier niveau d'analyse de ces corpus a été l'analyse multiparamétrique des signaux acoustiques. Les stimuli produits par le même locuteur ont ensuite fait l'objet d'évaluations perceptives.

3.6. *Matériel verbal et non verbal du corpus Sound Teacher*

Le corpus acté représente une durée d'enregistrement d'environ 2 heures (15 étiquettes émotionnelles par acteur en moyenne, exprimés sur les 10 phrases plus les 6 commandes utilisées dans la partie authentique).

L'enregistrement des 17 sujets dans la tâche prétexte d'apprentissage avec Sound Teacher représente une durée d'enregistrement de 15 heures. 6 heures sur 15 sont de la communication verbale, comprenant les commandes [ʁuʒ], [ʒon], [sabl], [vɛʁ], [bʁik], [paʒ sɥivãt] et les commentaires libres qui font l'objet d'une étude parallèle. Ces commentaires comprennent un large spectre d'attitudes, typiques des interactions homme-machine, dont l'analyse n'est pas décrite ici. Par ailleurs, ces commentaires présentent d'importantes variations dans le choix du lexique, des structures morpho-syntaxiques, et des structures de prosodie linguistique, qui sont les différentes composantes de l'expressivité. Pour 2 locuteurs enregistrés séparément, les commentaires ont été donnés par écrit. L'analyse des stratégies prosodiques et linguistiques intra et inter-locuteurs, l'étude contrastive des stratégies orales *vs.* textuelles, sont décrites par Terrier (2003).

Les 9 heures d'enregistrement restantes ne sont pas vides de signaux communicatifs : la partie non verbale de l'interaction contient de nombreux signaux et indices. En plus des événements « biologiques » (toux…) ou d'inconfort, les gestes faciaux des locuteurs, leurs postures, et leurs interjections (Ameka, 1992), *grunts* (Campbell *et al.*, 1992) et autres bruits de bouche tels que des raclements de gorge, des claquements de langue, expriment différents types d'information : changements émotionnels, attitudes et sentiments relatifs à l'interaction ou à la tâche qu'ils sont en train d'effectuer, états mentaux du locuteur, que nous regroupons dans le *feeling of thinking*, comme généralisation du *feeling of knowing* (Swerts et Krahmer, 2005). La quantité et la nature de telles informations sont des éléments éminemment pertinents dans l'interaction. Ne pas générer de telles expressions dans un agent conversationnel animé reviendrait à donner à l'interlocuteur humain pendant le tour de parole des messages partiellement faux, malformés et non écologiques ; ne pas reconnaître de telles informations en provenance de l'humain dans l'interaction personne-machine priverait le clone d'informations hautement pertinentes.

4. Analyse acoustique : de l'importance des contours

4.1. *Sélection des stimuli*

Pour cette première analyse, 136 stimuli produits par le même locuteur ont été extraits. Les stimuli actés sélectionnés ont été restreints aux énoncés comparables aux parties authentiques, excluant les phrases plus complexes également produites. 86 stimuli actés, porteurs d'expressions de satisfaction, surprise positive, concentration positive, inquiétude, anxiété, déception, joie tristesse, colère chaude, dégoût et peur, ainsi que d'expressions neutres, ont été sélectionnés, et 50 stimuli authentiques enregistrés à l'aide de Sound Teacher, étiquetés par le locuteur comme : confiance, concentration positive, joie/surprise, joie, concentration négative, déception/surprise, anxiété, anxiété/peur, lassitude et rien.

4.2. *Mesures acoustiques*

Un étiquetage phonétique expert, ainsi que l'étiquetage des mots, a été réalisé à l'aide du logiciel Praat (Boersma et Weenink) pour les parties actées et authentiques. De plus, les stimuli isolés et leur étiquetage ont été extraits du corpus brut à l'aide de scripts Praat.

Les contours de F0 ont été calculés pour chaque voyelle, à l'aide de l'éditeur prosodique EdiProso développé à l'ICP sous environnement Matlab, en se basant sur les bornes définies par les fichiers d'étiquettes. L'algorithme de calcul de F0 utilisé est basé sur une détection par seuil (fixé par défaut à 10 % de l'amplitude du signal). Les contours de F0 lissés, moyennés sur des fenêtres de 32 ms avec un décalage de 10 ms, ont été extraits à partir des valeurs calculées par cet algorithme. De plus, des contours normalisés sur 10 points ont également été extraits afin de permettre des comparaisons de formes de contours indépendamment de la durée de la voyelle.

Les durées des voyelles, extraites des fichiers d'étiquettes, ont été converties en durées relatives à la durée moyenne de la même voyelle dans le corpus afin de pouvoir effectuer des comparaisons intervoyelles. Les valeurs de l'attaque et les valeurs finales de F0 (mesurées avec un décalage de 10 % pour éviter les erreurs de calcul) pour chaque voyelle ont également été extraites, et utilisées pour calculer la déclinaison de F0 (différence attaque et finale), ainsi que la dynamique de F0, définie comme la différence entre les valeurs min et max. Toutes les valeurs de F0 ont été converties en demi-tons, la valeur nulle correspondant à la fréquence fondamentale moyenne du locuteur dans l'ensemble du corpus (96,8 Hz). La moyenne et l'écart type des valeurs d'attaque, de finale et de durée ont été calculées pour chaque étiquette émotionnelle. Calculer un contour-moyen (Aubergé, 1992) serait ici difficilement valide, car non seulement nous avons tout juste un nombre suffisant d'exemplaires pour que la moyenne soit statistiquement significative, mais surtout nous n'avons pas critère perceptif validé statistiquement (et évidemment pas morphologique puisque c'est ce que nous cherchons) qui étiquette l'intensité de l'émotion de chaque stimulus. Ainsi le contour-moyen représenterait une intensité non équilibrée statistiquement. C'est pourquoi nous avons opté pour une approche typiquement phonologique : nous avons sélectionné de « bons » exemplaires-prototypes (*i.e.* meilleurs scores d'identification perceptive, vérifiés par une écoute experte). Ces valeurs sont récapitulées dans le tableau 1.

4.3. *Résultats et discussion*

On peut tout d'abord remarquer que les contours étiquetés « neutre » (pour les émotions actées) et « rien » (pour les émotions authentiques) confirment l'hypothèse d'intonation minimale (segmentation/hiérarchisation et focalisation réduites) : en effet, l'attaque de ces deux expressions est au même niveau que la fréquence

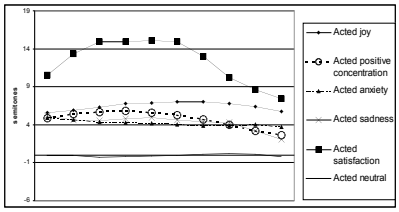
fondamentale moyenne du locuteur (que nous définissons dans notre modèle de prosodie (Aubergé, 2002) comme niveau de référence de F0, l'attaque étant considérée comme point d'ancrage des contours), la forme du contour est réduite à la ligne de déclinaison résultant de l'effort articulatoire sur de telles monosyllabes, qui correspond à la modalité déclarative.

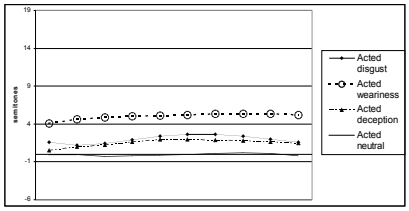
	Valence	Arousal	attaque F0	declin. F0	dyn. F0	durée norm. (%)
A anxiété	N	G	10	-1	1	-15,9
A déception	N	P	1	1	1,5	85,6
A dégoût	N	G	3	0	1	142
A peur	N	G	-4	6	6	14,5
A colère	P	G	15	3	3	29,2
A joie	P	G	11	0	1,5	16,2
A conc.pos.	P	P	10	-2	3	18,6
A surp.pos	P	G	-2	8	8	30,2
A lassitude	N	P	8	1	1	-2,9
A tritesse	N	G	10	-3	3	0,4
A satisfaction	P	P	21	-3	7	77,7
A inquiétude	N	P	0	11	11	17,9
A neutre	-	-	0	0	0,5	1,2
anxiété/peur	N	G	2	7	7	-6,6
confiance	P	P	3	-5	6	23,4
joie/surprise	P	G	-1	5	5	-12,6
lassitude	N	P	-3	2	2	-14,3
conc neg	N	P	2	3	3	-20,6
Rien	-	-	0	-2	2	-14,1
conc pos	P	P	1	-4	6	-1,3
Joie	P	G	1	5,5	5,5	-5,5
dec/surp	P	G	-1,5	7,5	7,5	-26,6
anxiété	N	P	1	7,5	7,5	-7,5

Tableau 1. *Valeurs caractéristiques des contours : attaque, déclinaison, dynamique, durée normalisée. Toutes les valeurs de F0 sont données en demi-tons. Les émotions actées sont signalées par 'A'. 'N'/'P' indique une valence négative/positive ; 'G'/'P' indique un grand/petit arousal*

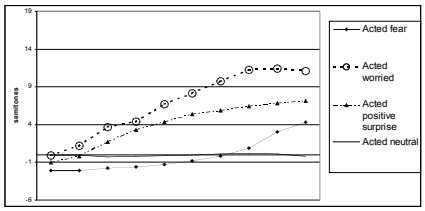
La dynamique moyenne des contours actés s'avère moins élevée (3,7 demi-tons) que celle des contours authentiques (5,2 demi-tons). Le niveau moyen de l'attaque

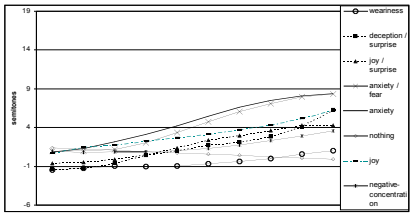
des contours actés est plus élevé (6,4 demi-tons) que celui des contours authentiques dont la moyenne est proche de 0 (avec toutefois d'importantes variations). En outre, la durée moyenne des voyelles est notablement plus élevée pour les expressions actées qu'authentiques (32 % vs. -8,6 %).





Figures 4. (gauche) et 5 (droite). *Expressions actées. Seule la satisfaction se détache, tandis que les contours de la joie, de l'anxiété, de la tristesse et de la concentration positive sont proches du neutre. Le dégoût, la lassitude et la déception ne présentent pas d'excursions particulières, mais ne suivent pas la ligne de déclinaison de base de l'expression neutre*





Figures 6. (gauche) et 7 (droite). *Peur, surprise et inquiétude actées présentent une forme similaire avec une proéminence finale. Les expressions authentiques de déception/surprise, joie/surprise, anxiété/peur, anxiété et joie d'une part, concentration négative et lassitude de l'autre, ont des formes similaires*

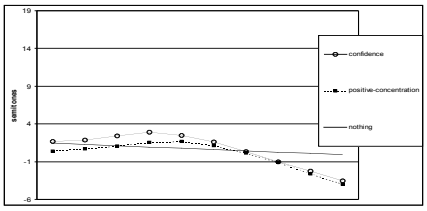


Figure 8. *Les expressions authentiques de la confiance et de la concentration positive ont des formes comparables avec une proéminence dans le premier tiers de la voyelle*

Parallèlement, nous avons symbolisé très sommairement les types de contours en 9 classes (/ ; ‾/‾ ; ‾/ ; /\ ; _/ ; ‾‾; ‾\ ; \/ ; _; \). De plus, des mesures quantitatives des paramètres acoustiques décrits par Scherer *et al*., (2003) ont été effectuées (moyenne, écart type, étendue, percentiles, min/max et jitter pour F0, ainsi que les 11 paramètres spectraux décrits dans cette étude), de même que des mesures du paramètre de source glottique NAQ (Alku *et al.*, 2002). Le seul effet significatif (ANOVA, p = .01) ressortant des analyses statistiques effectuées est l'effet du contour /\ sur NAQ, sur le jitter et sur la pente spectrale. Toutefois le choix des classes symboliques de contours à partir de leur forme n'est pas univoque, et le choix de ces classes induit très probablement une perte d'information : le symbolisme devrait inclure certains indices pouvant être observés sur les figures présentées ci-dessus. En particulier, la pertinence de la localisation et du seuil du glissando, validé en psychoacoustique (Rossi, 1999) mais non pertinent pour la prosodie linguistique, pourrait être évalué pour les valeurs de prosodie émotionnelle, notamment quand le timing n'est pas lié aux unités linguistiques.

4.4. *En résumé*

Bänziger *et al*., (2003) ont obtenu des résultats perceptifs montrant que seul le niveau général de F0 est pertinent pour la perception des expressions émotionnelles, et confirment que la hauteur moyenne peut être liée aux valeurs d'activation et d'arousal. Quoique ces caractéristiques soient vraisemblablement les plus pertinentes, nous suggérons que des indices extra-linguistiques plus fins véhiculés par la forme des contours ont pu être masqués dans cette expérience par les contours linguistiques porteurs, bien que ces derniers soient supposés identiques pour les phrases longues choisies. Il serait donc intéressant de répliquer une étude similaire sur des unités linguistiques minimales, en gelant également les variations pragmatiques. Les contours de F0 stylisés et paramétrisés par un ensemble de valeurs d'ancrage qui sont présentés ici permettent d'observer différentes classes de contours lorsqu'on les compare aux contours de référence (pour la parole actée et authentique). Bien que le lien ne soit pas établi ici entre l'étiquetage des expressions et la morphologie des contours, ces contours sont suffisamment simples pour tester différents types de stylisation dans le cadre d'expériences perceptives. Dans cette optique il sera nécessaire de procéder à des prétests perceptifs pour déterminer lesquels des indices de ces contours (par exemple, l'emplacement du glissando) sont sensibles aux différentes étiquettes. Préalablement à cette validation perceptive, des premières conclusions peuvent être tirées de cette étude : tout d'abord il existe bien un contour « sans valeur émotionnelle », aussi bien pour les expressions actées qu'authentiques, et que la forme de ce contour correspond exactement à celle attendue sur de telles unités linguistiques minimales, les autres variations étant gelées par ailleurs ; de plus, il apparaît également clairement que la forme des contours varie avec les émotions exprimées.

5. Evaluation perceptive

Une fois les analyses acoustiques menées à bien, des mesures perceptives ont été réalisées, afin dans un premier temps de valider les enregistrements, puis par la suite de relier la perception des émotions avec les variations individuelles des paramètres acoustiques.

Dans ce but, deux tests de perception ont pour l'instant été effectués. Le premier porte sur les stimuli enregistrés durant la phase de production actée décrite ci-dessus. Le test consiste donc à valider non seulement la performance des acteurs, mais aussi les choix des annotations qui ont été utilisées pour reproduire les émotions durant la phase actée. Le second test de perception porte sur un sous-ensemble des stimuli validés lors du premier test, modifiés par analyse-synthèse afin de faire varier séparément les paramètres prosodiques. Cette procédure permet de mesurer le poids relatif de chaque paramètre sur la perception des expressions émotionnelles.

5.1. *Test de validation du corpus de parole émotionnelle*

Ainsi que nous venons de l'expliquer, l'un des buts principaux de ce test de perception est de valider les annotations émotionnelles données par les locuteurs. Ceci est nécessaire avant de pouvoir aller plus loin dans notre étude, car nous ne connaissons pas *a priori* la pertinence cognitive de ces annotations, non plus que la pertinence des performances de l'acteur au regard de chacune de ces annotations. Cela revient à répondre aux questions suivantes : les annotations utilisées correspondent-elles aux expressions réalisées par l'acteur ? Deux expressions différentes renvoient-elles systématiquement à des annotations différentes ou certaines sont-elles confondues, et inversement chaque annotation renvoie-t-elle à une expression différente ou certaines peuvent-elles être confondues ?

5.1.1. *Sélection des stimuli*

Les stimuli utilisés pour le test de perception sont ceux enregistrés par un locuteur masculin, acteur. Il s'agit ici uniquement des stimuli obtenus lors des performances actées. En effet, le but de ces premiers tests perceptifs est en partie la validation des annotations utilisées et de leur pertinence : il n'est donc pas envisageable d'utiliser directement les stimuli authentiques, pour lesquels on ne connaît pas *a priori* l'expression véhiculée. Le choix du locuteur est en partie dicté par les analyses acoustiques, effectuées préalablement sur les stimuli produits par ce locuteur.

La sélection des enregistrements a été faite par deux experts qui se sont concertés pour noter sur une échelle allant de 1 (mauvaise performance) à 4 (très bonne performance) l'ensemble des productions du locuteur. Seuls les stimuli ayant reçu une note de 3 ou 4 ont été sélectionnés, et parmi ceux-ci un sous-ensemble a été

sélectionné, avec comme critères une variation systématique de la longueur (choix de stimuli de 1, 3, 5 et 7 syllabes), la présence de toutes les émotions actées (au nombre de 14 : amusement, colère, anxiété, déception, dégoût, attente, peur, joie, neutre, résignation, tristesse, satisfaction, surprise, inquiétude) représentées pour chaque longueur de stimulus, ainsi qu'un stimulus identique (« page suivante ») pour chaque expression, pour un total de 70 stimuli différents.

5.1.2. *Protocole expérimental*

26 sujets âgés de 19 à 45 ans (25 en moyenne) ont participé à cette expérience. La tâche consistait à juger l'intensité perçue pour chacune de ces 14 émotions sur une échelle de 0 à 10. Pour cela, les sujets entendaient une fois chaque stimulus dans un ordre aléatoire, et devaient positionner un curseur sur une échelle de 0 à 10 pour chacune des 14 émotions, plusieurs émotions pouvant être exprimées par un même stimulus. Les stimuli étaient d'abord présentés en audio seul, puis dans une seconde condition expérimentale, en audio-vidéo. Ceci a été réalisé afin d'avoir une condition de contrôle lorsque les sujets ne parviennent pas à reconnaître une émotion grâce à la modalité audio seule.

5.1.3. *Résultats*

La valeur élevée de l'alpha de Cronbach ($\alpha = 0{,}95$) traduit une bonne cohérence des réponses entre les 26 sujets, indiquant une stratégie homogène dans la tâche d'identification qui leur est proposée. Des matrices de confusion ont été établies afin d'analyser les scores de reconnaissance ainsi que les éventuelles dispersions sur d'autres émotions. Ces matrices de confusions ont été analysées telles quelles, grâce à un traitement en grande partie qualitatif. En effet, ces données ne permettent pas un traitement statistique poussé car elles sont trop complexes pour rentrer dans le cadre d'une analyse de la variance. Par contre, ce choix permet de conserver toute l'information sur la distribution des réponses dans les différentes catégories proposées aux sujets. Cela nous a semblé primordial, car le but premier de cette expérience est de savoir ce que les auditeurs perçoivent vraiment, et non pas de savoir simplement s'ils perçoivent correctement l'expression que le locuteur est censé avoir réalisée : nous cherchons à mesurer ce qui est transmis par ces stimuli plutôt qu'à évaluer la performance de l'acteur. Cette analyse permet donc de mettre en lumière les comportements suivants :

– en condition audio seul, le dégoût est reconnu avec un fort effet auditeur, alors que les résultats de la condition audiovisuelle sont tranchés, ce qui confirme les résultats rapportés par Scherer (2003) et Juslin & Laukka (2003) à ce sujet ;

– certaines catégories ne sont pas utilisées : « *attente* », reconnue comme « *neutre* » ; « *déception* », reconnue comme « *résignation* ». La condition audiovisuelle n'améliore pas ces résultats ;

– la catégorie « neutre » est la catégorie utilisée pour les expressions non reconnues : « *attente* », « *joie* » (la joie est actée de manière peu activée par ce locuteur) et « *résignation* » ;

– certaines expressions sont confondues entre elles : il s'agit de « *anxiété, inquiétude, peur* » et de « *amusement, joie, satisfaction* » ;

– les expressions les mieux reconnues sont « *amusement* », « *anxiété* », « *colère* », « *neutre* », « *satisfaction* » et « *surprise* ».

Enfin, nous avons voulu vérifier si les distributions des réponses des sujets sur les différentes émotions proposées variaient avec l'accroissement de longueur des stimuli. Pour cela, tous les résultats des stimuli d'une longueur donnée ont été regroupés. La corrélation de ces distributions moyennes classées par longueur de stimuli (1, 3, 5 et 7 syllabes) a été calculée. Ces corrélations sont toutes significatives (à $p < 0,05$), ce qui montre le peu d'influence de la longueur sur les traitements perceptifs de ce type d'informations expressives.

5.2. *Evaluation des poids relatifs des paramètres prosodiques*

Afin d'évaluer le poids de chaque paramètre prosodique dans la morphologie vocale des émotions, et donc dans la perception de leurs expressions, il est nécessaire d'évaluer comment des stimuli porteurs de ces expressions sont perçus lorsqu'on agit séparément sur ces paramètres. Étant donné que les variations de ces paramètres résultent d'un contrôle global du conduit vocal, il semble impossible de recueillir naturellement de tels stimuli, quand bien même ils seraient produits par des acteurs entraînés spécifiquement pour cette tâche. En revanche une autre méthode, que nous avons retenue, consiste à procéder à une analyse acoustique de stimuli de référence avant de synthétiser de nouveaux stimuli à partir de tout ou partie des paramètres analysés, en fonction des hypothèses à tester. Enfin, une évaluation perceptive de ces stimuli permet la validation de ces hypothèses. Dans le cadre de l'étude des expressions émotionnelles, une telle méthode a notamment été utilisée pour étudier la pertinence de la qualité de voix (Gobl *et al.*, 2000), pour tester l'influence relative de la fréquence fondamentale et de l'intensité *vs.* la qualité de voix (Bulut *et al.*, 2002), ou encore pour évaluer la pertinence des mesures multiparamétriques effectuées sur les signaux de parole (Bänziger *et al.*, 2003).

5.2.1. *Génération des stimuli synthétiques*

Nous avons sélectionné 10 stimuli monosyllabiques de référence extraits du corpus E-Wiz/Sound Teacher et constituant un sous-ensemble des 70 stimuli actés précédemment évalués dans l'étude perceptive décrite ci-dessus. Ces 10 stimuli expriment sur les mots monosyllabiques français [ruʒ] et [sabl] les états émotionnels suivants : anxiété, déception, dégoût, inquiétude, joie, résignation, satisfaction et tristesse. De plus, des expressions neutres de ces 2 mots ont été sélectionnés comme référence pour la comparaison des contours multiparamétriques. La pertinence d'une expression neutre pour le locuteur sélectionné est validée par la présence dans ses productions authentiques d'expressions étiquetées comme « rien » par le locuteur lui-même.

Les contours multiparamétriques ont été analysés et utilisés en resynthèse à l'aide du logiciel Praat (Boersma et Weenink) en suivant une procédure semi-automatique, qui consistait à styliser les contours de F0 et d'intensité extraits d'un stimulus source, d'appliquer tout ou partie de ces contours à un stimulus cible et de synthétiser un nouveau stimulus combinant des propriétés des stimuli source et cible en utilisant le module de synthèse basé sur TD-PSOLA de Praat.

Pour chacun des stimuli originaux porteurs d'une expression émotionnelle, 5 stimuli distincts ont ainsi été générés : i) un stimulus de contrôle étiqueté *resynthèse complète*, construit en appliquant les contours stylisés de F0 et d'intensité du stimulus source à lui-même et destiné à évaluer d'éventuels artefacts dus au processus de resynthèse ; ii) un stimulus étiqueté *F0 seule*, construit en appliquant le contour stylisé de F0 du stimulus source au stimulus porteur d'une expression neutre correspondant au même mot ; iii) un stimulus étiqueté *intensité seule* obtenu en appliquant le contour d'intensité du stimulus source à l'expression neutre correspondante ; iv) un stimulus étiqueté *F0 et intensité* construit en appliquant les contours de F0 et d'intensité à l'expression neutre correspondante ; v) un stimulus étiqueté *qualité de voix et durée*. Cette dernière condition a été obtenue en appliquant les contours de F0 et d'intensité de l'expression neutre au stimulus source. Ainsi seuls les phénomènes de durée et de qualité de voix du stimulus source subsistent, tandis que ses variations spécifiques de F0 et d'intensité sont neutralisées. En complément des 40 stimuli générés à partir des 8 expressions émotionnelles sélectionnées, un stimulus en condition *resynthèse complète* a été généré pour chacune des 2 expressions neutres.

Un inconvénient de la méthode utilisée pour la génération des stimuli en condition *qualité de voix et durée* est que les phénomènes de qualité de voix et de durée ne peuvent pas être traités séparément. Toutefois, étant donné que l'une de nos hypothèses est que les expressions directes des émotions ne sont pas régies par le domaine temporel des fonctions linguistiques de la prosodie (Aubergé, 2002), nous avons décidé de ne pas nous baser sur un modèle linguistique de durée potentiellement inadapté aux phénomènes étudiés. De plus, en l'absence de méthode robuste pour suivre un paramètre de qualité de voix, nous n'avons pas été en mesure de traiter la qualité de voix de façon fiable dans cette étude. Le quotient d'amplitude normalisé NAQ (Alku *et al.,* 2002) semblait être un bon candidat, validé pour la caractérisation de très grands échantillons de parole (Campbell *et al.*, 2003) ; toutefois, nous avons pu montrer un effet intrinsèque majeur de la qualité de la voyelle dans le calcul de NAQ qui nous a forcé à rejeter NAQ comme paramètre pertinent pour le suivi de contours de qualité de voix (Rossato *et al.*, 2004),

5.2.2. *Evaluation perceptive des stimuli générés*

Les 42 stimuli générés ont été notés par 40 juges de langue maternelle française (6 hommes et 34 femmes, d'âge moyen 23,3 ans) au laboratoire en chambre sourde, chaque stimulus étant présenté 3 fois. Les stimuli étaient présentés dans un ordre aléatoire différent pour chaque juge et contrôlé pour éviter deux présentations

successives du même stimulus. Les sujets avaient pour instruction de sélectionner l'une des 8 étiquettes émotionnelles proposées (anxiété, déception, dégoût, inquiétude, joie, résignation, satisfaction ou tristesse) ou l'étiquette neutre lorsque le stimulus était perçu comme n'exprimant aucune émotion, et de noter l'intensité émotionnelle perçue entre 1 et 10. La valeur élevée de l'alpha de Cronbach ($\alpha = 0{,}954$) indique une bonne cohérence entre les réponses données par les différents juges.

Les résultats de l'évaluation perceptive ont été traités séparément pour chaque condition de resynthèse. Étant donné que les intensités émotionnelles attribuées n'apportent que peu d'information, seuls les scores d'identification compilés en 5 matrices de confusion sont présentés ici. En raison du nombre important de confusions, il est difficile de tirer des conclusions directement à partir des matrices de confusions. Toutefois certaines tendance apparaissent : ainsi l'expression de la satisfaction (reconnue à 24,2 % en condition de contrôle) obtient des scores presque aussi élevés en conditions *F0 seule* (21,7 %) et *F0 et intensité* (22,5 %). De plus, l'expression de l'anxiété (reconnue à 57,5 % en condition de contrôle) obtient également un taux d'identification élevé en condition *qualité de voix et durée* (53,3 %). Afin de prendre en compte les principales confusions et faire ressortir plus clairement les principales tendances, certaines catégories ont été regroupées : la joie avec la satisfaction, l'anxiété avec l'inquiétude, et la tristesse avec la déception et la résignation. En revanche, dégoût et neutre restent des catégories distinctes.

De nouvelles matrices de confusion ont été calculées afin de prendre en compte ces regroupements, les scores étant pondérés par le nombre d'étiquettes initiales regroupées pour former la catégorie. Le test du khi-deux indique que ces distributions sont toutes significativement différentes du hasard ($p <= 0{,}01$). De plus, l'essentiel de l'information se trouvant sur les diagonales des matrices de confusions ainsi regroupées, les données ont été converties en bonnes/mauvaises réponses et normalisées afin de pouvoir effectuer une ANOVA mesures répétées. Cette analyse de variance montre un effet de la condition de synthèse, de l'expression, et de l'interaction condition*expression, et permet de tester la significativité des contrastes. Les différences mentionnées ci-dessous sont significatives à $p = 0{,}01$ sauf mention contraire. La figure 9 présente un ensemble de diagrammes illustrant les résultats après regroupement pour chaque condition de resynthèse, ainsi qu'une comparaison des scores d'identification obtenus dans les différentes conditions. Les positions des étiquettes dans les diagrammes sont fonction du taux d'identification et de l'attractivité de la catégorie correspondante. L'attractivité a été définie pour une étiquette donnée comme la somme des réponses attribuées à tort à cette étiquette, normalisée par le score théorique maximum. De plus, les flèches entre étiquettes indiquent les confusions supérieures au hasard, leur épaisseur étant proportionnelle au taux de confusion.

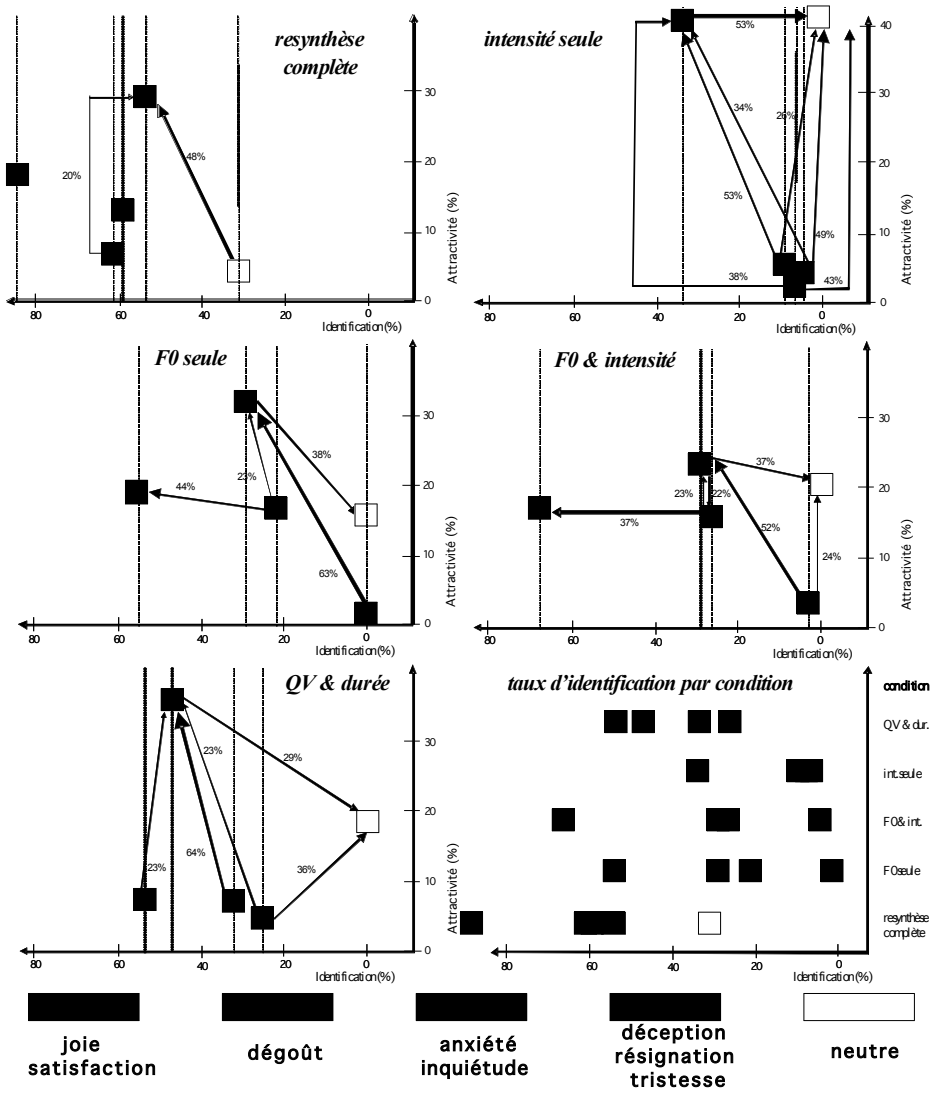


Figure 9. *Taux d'identification et confusions après regroupement de catégories pour chaque condition de resynthèse. La figure en bas à droite compare les scores d'identification entre conditions*

En condition *intensité seule*, la plupart des expressions ont obtenu des taux d'identification inférieurs à 10 % et non significativement différents, ces expressions étant perçues comme neutre, à l'exception des expressions de tristesse, résignation et déception identifiées à 35,6 %. Toutefois cette différence peut s'expliquer par le fait

que les sujets percevant l'une de ces expressions dans cette condition ont également largement confondu les expressions neutres avec la tristesse, la résignation ou la déception en condition de contrôle.

En conditions *F0 seule* et *F0 et intensité*, les expressions de joie et satisfaction ont été bien reconnues (respectivement à 56,3 % et 67,5 % vs. 85,4 % en condition de contrôle), tandis que anxiété et inquiétude ainsi que tristesse, résignation et déception n'ont été reconnues qu'avec un score légèrement supérieur au hasard, et que le dégoût a été systématiquement confondu. La comparaison des scores en condition *F0 seule* vs. *F0* et intensité indique que, bien que l'intensité seule soit insuffisante pour identifier ces expressions, l'ajout de l'information d'intensité permet d'améliorer significativement l'identification des expressions d'anxiété et d'inquiétude, ainsi que de joie et de satisfaction ($p = 0,05$ pour ces dernières).

Enfin, en condition *qualité de voix et durée*, tristesse, résignation et déception (55,8 %) ainsi qu'anxiété et inquiétude (47,2 %) ont été les expressions les mieux reconnues, et le dégoût (34,2 %) a obtenu un meilleur score que la joie (24,6 %). Étant donné que les autres paramètres n'apparaissent pas comme des indices fiables pour l'identification du dégoût, il est assez surprenant que le dégoût dans cette condition soit reconnu avec un score qui ne représente que 55 % du score obtenu en condition de contrôle. L'expression du dégoût dans sa modalité audio semble donc moins robuste que les autres expressions testées, ce qui pourrait expliquer qu'il soit souvent décrit dans la littérature comme mal identifié (voir Juslin et Laukka, (2003) pour un état de l'art).

5.2.3. *En résumé*

Cette étude a été menée sur des mots monosyllabiques supposés n'exprimer que des émotions directes, l'expression des autres affects (attitudes et expressivité) étant gelée. Les stimuli qui ont été synthétisés en faisant varier un par un les paramètres prosodiques constituent des artefacts. Toutefois, les auditeurs ont pu identifier des traits saillants de cette prosodie chimérique, notamment à partir des variations de F0 pour les émotions positives, et de la qualité de voix et de la durée pour les émotions négatives. Ces résultats nous informent sur les paramètres les plus pertinents dans une optique de modélisation prosodique pour la synthèse d'expressions émotionnelles. Les contours d'intensité utilisés seuls se sont révélés insuffisamment informatifs pour discriminer les stimuli synthétiques des expressions neutres, mais ont permis de renforcer significativement l'identification de certaines expressions, ce qui met l'accent sur la nécessité de prendre en compte les paramètres moins prégnants afin de générer des expressions qui ne soient plus des chimères. De plus, ces résultats soulignent l'importance des contours, au-delà des valeurs globales des paramètres, comme une dimension pertinente pour la génération des expressions.

6. Conclusion

Les fonctions affectives encapsulées en trois traitements, sont une composante indissociable des fonctions communicatives, si l'on s'appuie sur les hypothèses de prise de décisions (Damasio, 1994) et d'évaluation (Lazarus 1991; Scherer, 2001) et de préparation à l'action (Frijda, 1987). C-Clone est une architecture qui propose d'intégrer les traitements linguistiques et prosodiques dans le même processus coopératif. C-Clone est caractérisé par un « fonctionnalisme corporéisé» ; les fonctions communicatives régissent le système à trois niveaux de granularité : interfonctionnel, intrafonctionnel et modulaire. Chaque fonction est définie à travers son implémentation « physique » et résulte de l'intégration entre un contrôle systémique et la morphologie du module physique qui véhicule l'information communicative.

La plateforme E-Wiz a été construite comme un outil ouvert et gratuit pour la conception spécifique de protocoles de capture d'expressions authentiques en interaction personne-machine, couvrant les niveaux d'affects qui sous-tendent l'architecture de C-Clone. Un corpus d'expressions authentiques d'émotions, d'attitudes et d'expressivité, recueilli à l'aide d'E-Wiz, dans le scénario Sound Teacher, a permis l'étude spécifique des expressions émotionnelles involontairement contrôlées dans la voix. Le protocole d'auto-annotation mis en place a éludé le problème du choix théorique de la nature cognitive des émotions (dimensions *vs.* catégories) et s'est avéré une monnaie d'échange pertinente pour les auditeurs dans les tâches de validation/évaluation. L'analogie entre les annotations des auditeurs, montre des réactions analogues des sujets au scénario et permet d'étudier à la fois les variations intra et interlocuteurs. Le spectre des émotions annotées est typique de situations d'interaction personne-machine, incluant des émotions positives à négatives, actives à passives. Nous avons pu ainsi intervenir dans la discussion initiée par Fonagy (1986) et Delattre (1966), et prolongée par Ladd *et al.*, (1984), Scherer (2001), Bänziger *et al*., (2003), Mozziconnacci (2001) et Gobl et Ní Chasaide (2000) sur 1. la morphologie de la prosodie émotionnelle comme formes *vs.* valeurs globales, et 2. l'attribution de certains paramètres acoustiques à la prosodie émotionnelle *vs*. linguistique. Si la gradience peut être une caractérisation suffisante en reconnaissance des formes prosodiques émotionnelles – comme proposé par Scherer *et al*., (2003)), également utilisée dans des approches statistiques par Batliner (2003) – la génération des formes dans toute leur caractéristique globale est nécessaire en génération, si l'on veut éviter un effet « chimère », le résultat principal de nos analyses étant qu'aucun paramètre n'est dédié à la parole émotionnelle, et qu'il n'existe pas non plus de paramètre ne portant aucune information émotionnelle.

La modélisation des expressions en contours gradients qui intègrent dans le même processus morphologique les indices émotionnels et la variabilité de l'intensité sur ces indices nous ont conduit à généraliser le modèle général de morphologie de l'ensemble des fonctions communicatives de la prosodie. La

génération des expressions émotionnelles constitue le niveau primaire du modèle de la prosodie mis en œuvre dans C-Clone pour la prosodie linguistique et attitudinale (Aubergé, 1992 ; Aubergé, 2002). Il est basé sur un principe de superposition de contours prototypiques contrôlés par les fonctions communicatives. Pour implémenter une génération pertinente de ces expressions émotionnelles, la dimension de la qualité de voix doit être modélisée en un ou plusieurs paramètres acoustiques (Rossato *et al*., 2004). Des algorithmes de codage de la qualité de voix sont en cours d'évaluation à travers une coopération avec France Télécom R&D (Audibert *et al*., à paraître).

Remerciements

Ce travail a débuté dans le programme « Expressive Speech Processing » du Crest/Japan Science Technology de l'équipe de Nick Campbell, avec le soutien du projet « Communication Expressive » du PPF Pegasus des universités grenobloises, et en partie dans le cadre d'une coopération par Contrat de Recherche Externalisée avec France Télécom R&D. Nous tenons à remercier Nick Campbell pour les discussions constructives menées au sein du CREST, Christophe Savariaux et Alain Arnal, responsables du bloc expérimental de l'ICP, sans qui les corpus E-Wiz n'auraient pas été réalisables.

7. Bibliographie

Alku P., Bäckström T., Vilkman E., "Normalized amplitude quotient for parameterization of the glottal flow", *Journal of the Acoustic Society of America*, vol. 112, (2), 2002, p. 701-710.

Ameka F., "Interjections: The universal yet neglected part of speech", *Special issue of Journal of Pragmatics*, vol. 18, (2), 1992, p. 110-118.

Amir N., Ron S., "Toward an Automatic Classification of Emotions in Speech", *ICSLP*, Sydney, Australie, 1998, p. 29-33.

Aubergé V., "Developing a structured lexicon for synthesis of prosody", *Talking machine*, Bailly & Benoit Eds, Elsevier, 1992.

Aubergé V., Grépillat T., Rilliard A., "Can we perceive attitudes before the end of sentences? The gating paradigm for prosodic contours", *Eurospeech*, Rhodes, Grèce, 1997, p. 871-874.

Aubergé V., "A Gestalt morphology of prosody directed by functions: the example of a step by step model developed at ICP", *Speech Prosody*, Aix-en-Provence, 2002, France, p. 151-155.

Aubergé V., Cathiard M., "Can we hear the prosody of smile ?", Special Issue *Emotional Speech*, *Speech Communication Review*, 40, 2003, p. 87-97.

Aubergé V., Audibert N., Rilliard A., "E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab.", *LREC*, Lisbon, Portugal, 2004, p. 179-182.

Aubergé V., Rilliard A., "More than pointing with the prosodic focus: The Valence-Intensity-Domain (VID) model", *International Conference on Speech Prosody* 2006, à paraître.

Audibert N., Vincent D., Aubergé V., Rosec O. "The prosodic dimensions of emotion in speech: the relative weights of parameters", *InterSpeech* 2006 à paraître.

Bänziger T., Morel M., Scherer K. R., "Is there an emotion signature in intonational patterns? And can it be used in synthesis?", *Eurospeech*, Geneva, Suisse, 2003, p. 1641-1644.

Batliner A., Fischer K., Huber R., Spilker J., Nöth E., "How to Find Trouble in Communication", *Speech Communication*, vol. 40, 2003, p. 117-143.

Boersma P, Weenink D., Praat, doing phonetics by computer, Institute of Phonetic Sciences, University of Amsterdam, Pays-Bas, 1992-2004.

Bulut M., Narayanan S. S., Syrdal A. K., "Expressive speech synthesis using a concatenative synthesiser", *2002 7th ICSLP*, Denver, Colorado, USA.

Campbell N., "Databases of Emotional Speech", *2000 ISCA Workshop on Speech and Emotions*, Newcastle, Irlande du Nord, p. 34-38.

Campbell N., Mokhtari P., "Voice Quality: the 4th Prosodic Dimension", *2003 15th ICPhS*, Barcelona, Espagne, p. 2417-2420.

Clément J., Structure des représentations prosodiques, Développement normal et pathologique du traitement de la prosodie, Thèse de doctorat, Université Paris V, 1999.

Cornelius R. R., *The science of emotion. Research and tradition in the psychology of emotion*, Upper Saddle River (NJ): Prentice-Hall, 1996.

Damasio A. R., *Descartes' error. Emotion, reason, and the human brain*, A. Grosset/Putnam Books, 1994.

Delattre P., « Les dix intonations de base du français », *The French Review*, vol. 40, (3), 1966, p. 1-14.

Douglas-Cowie E., Cowie R., Schröder M., "A new emotion database: considerations, sources and scope", *2000 ISCA Workshop on Speech and Emotions*, Newcastle, Irlande du Nord, p. 39-44.

Fonagy I., *La vive voix*, Paris, Payot, 1983.

Frijda N. H., "Emotions, Cognitive structures and Action tendency", *Cognition and Emotion*, vol. 1, 1987, p. 115-143.

Gobl C., Ní Chasaide A., "Testing affective correlates of voice quality through analysis and resynthesis", *2000 ISCA Workshop on Speech and Emotions*, Newcastle, Irlande du Nord, p. 178-183.

Iida A., A Study on Corpus-based Speech Synthesis with Emotion, Thèse de doctorat, Keio University, 2002.

Johnstone T., Scherer K. R., "The effects of emotions on voice quality", *1999 14th ICPhS*, San Francisco, USA, p. 2029-2032.

Juslin P. N., Laukka P., "Communication of emotions in vocal expression and music performance: Different channels, same code?", *Psychological Bulletin*, vol. 129, (5), 2003, p. 770-814.

Kaiser S., Wehrle T., "Emotion research and AI: some theoretical and technical issues", *Geneva Studies in Emotion and Communication*, vol. 8, (2), 1994, p. 1-16.

Ladd D. R., Silverman K., Tolkmitt F., Bergmann G., Scherer K. R., "Evidence for the independent function of intonation contour type, voice quality and f0 range in signalling speaker affect", *Journal of the Acoustic Society of America*, vol. 78, (2), 1985, p. 435-444.

Lazarus R. S., "Progress on a cognitive motivational-relational theory of emotion", *American Psychologist*, 46, (8), 1991, p. 819-834.

Leinonen L., Hiltunen M. L., "Expression of emotional motivational connotations with one-word utterance", *Journal of the Acoustic Society of America*, vol. 102, (3), 1997, p. 1853-1863.

Léon P. R., *Précis de phonostylistique : parole et expressivité*, Nathan université, 1993.

MacNeil D., *Hand and Mind*, University of Chicago Press, 1992.

Martin J.C., Abrilian S., Devillers L., "Annotating multimodal behaviors occurring during non basic emotions", *Affective Computing and Intelligent Interaction*, Beijing, 2005, p. 550-557.

Mokhtari P., Campbell N., "Automatic Detection of Acoustic Centres of Reliability for Tagging Paralinguistic Information in Expressive Speech", *2002 3rd LREC*, Las Palmas, Espagne, 2002, p. 2015-2018.

Morlec Y., Bailly G., Aubergé V., "Generating prosodic attitudes in French: data, model and evaluation", *Speech Communication*, vol. 33, (4), 2001, p. 357-371.

Mozziconacci S., Speech Variability and Emotion: Production and Perception, Thèse de doctorat, Eindhoven University, Pays-Bas, 1998.

Rilliard A., Aubergé V., "Prosody evaluation as a diagnostic process: Subjective vs. objective measurement", *International Journal of Speech Technology*, Kluwer Academic Publishers, 2003, p. 409-418.

Rossato S., Audibert N., Aubergé V., "Emotional Voice Measurement: A Comparison of Articulatory-EGG and Acoustic-Amplitude Parameters", *2004 Speech Prosody*, Nara, Japan, p. 749-752.

Rossi M., *Intonation: past, present and future*, A. Botonis (ed), Elsevier, 1999.

Scherer K. R., Ladd D.R., Silverman K. E. A., "Vocal cues to speaker affect: testing two models", *Journal of the Acoustic Society of America*, vol. 76, (5), 1984, p. 1346-1356.

Scherer K. R., "Appraisal considered as a process of multi-level sequential checking", In Scherer K. R., Schorr A., & Johnstone T. (Eds.), *Appraisal processes in emotion: Theory, Methods, Research*, Oxford University Press, 2001, p. 92-120.

Scherer K. R., Johnstone T., Klasmeyer G., "Vocal Expression of Emotion", In Davidson R. J., Scherer K. R., Goldsmith H. H. (Eds), *Handbook of Affective Sciences*, 2003, p. 433-456.

Shochi T., Aubergé V., Rilliard A., "How French can hear Japanese: the kyoshuku politeness misunderstanding", *Humaine NoE Workshop*, Paris, 2005.

Swerts M., Krahmer E., "Audiovisual prosody and feeling of knowing", *Journal of Memory and Language*, 2005 à paraître.

Terrier M., Stratégies expressives écrit versus oral: quels degrés de liberté de l'oral ?, Rapport de Master 2 en Sciences Cognitives, Institut National Polytechnique de Grenoble, 2003.

Wichmann A., "Attitudinal Intonation and the Inferential Process", *2002 Speech Prosody*, Aix-en-provence, France, p. 11-15.

Williams C. E., Stevens K. N., "Emotions and speech: some acoustical correlates", *Journal of the Acoustic Society of America*, vol. 52, 4, (2), 1972, p. 1238-1250.