

Université Stendhal – Grenoble III

T.E.R. de Maîtrise Sciences du Langage
Mention Industries de la Langue

L'expression vocale de l'amusement : premières expériences audiovisuelles



Institut de la Communication Parlée

Marc SCHRÖDER

Responsable : Véronique AUBERGÉ

Juin 1998

*Puisse ce travail apporter une poussière
à la compréhension de l'être humain
et ainsi contribuer, aussi infiniment peu que ce soit,
à la paix et au bonheur pour tous.*

REMERCIEMENTS

Je voudrais remercier en premier lieu ma directrice de T.E.R., Véronique Aubergé, qui m'a permis de travailler sur le sujet de l'expression de l'émotion. Je la remercie du temps qu'elle m'a consacré ; des discussions sur des sujets pas toujours faciles, mais qui m'ont éclairci des aspects intéressants de son point de vue ; des conseils en matière de style, et de sa relecture suivie des parties achevées du manuscrit. Sa gestion du temps, généralement très différente de la mienne, m'a rendu plus flexible.

Merci aussi à Christian Abry, qui nous a conseillé à un moment crucial : c'est lui qui nous a proposé de travailler sur le sourire, en nous indiquant les avantages et les possibilités impliquées par ce choix. Sans lui, ce travail n'aurait pas pris la même voie. Merci pour cela.

Un remerciement particulier est réservé pour Alain Arnal, qui nous a apporté de l'aide précieuse tout au long de la partie pratique de ce travail : de la préparation du matériel d'enregistrement, en passant par le coloriage des dessins utilisés pendant l'enregistrement du corpus, maintes explications techniques, le soutien infatigable à la recherche d'une solution pour la numérisation, jusqu'à la programmation en Hypercard d'un outil pour mesurer la configuration faciale – tout ceci a fait d'Alain Arnal un soutien inestimable pour ce T.E.R. Merci, Alain, aussi pour ta bonne humeur contagieuse.

Ensuite, il m'importe de remercier tous ceux qui m'ont apporté de l'aide, notamment Marie Cathiard, pour des conseils méthodologiques, et qui a rendu possible l'analyse ANOVA des résultats du test de perception, et Albert Rilliard, qui m'a conseillé en questions de méthodologie et d'utilisation de Macintosh patiemment pendant toute l'année.

Merci enfin à Amelyne, qui a relu le manuscrit et corrigé des fautes de français, qui a partagé ma joie et ma frustration, et qui a toujours trouvé des mots encourageants quand il en fallait.

TABLE DES MATIERES

INTRODUCTION	1
LES ÉMOTIONS : FONCTIONS ET PHYSIOLOGIE — UN SURVOL BIBLIOGRAPHIQUE	3
1. QUELQUES NOTIONS THÉORIQUES SUR LES ÉMOTIONS	3
1.1. <i>Les fonctions des émotions</i>	3
1.2. <i>Un survol des différents types de théories</i>	4
1.3. <i>Le modèle des processus composants de Scherer</i>	5
2. NEUROPHYSIOLOGIE DES ÉMOTIONS : DAMASIO 1994	6
2.1. <i>Son approche générale</i>	6
2.2. <i>Les émotions selon Damasio</i>	6
2.2.1. Les émotions primaires.....	7
2.2.2. Les émotions secondaires.....	7
2.3. <i>La spécificité des mécanismes neuraux sous-tendant les émotions : l'exemple du sourire</i>	8
2.3.1. Emotions et l'hémisphère droit	8
2.3.2. La différence entre le sourire spontané et le sourire volontaire	8
2.4. <i>La perception des émotions</i>	9
2.4.1. Tromper le cerveau – la perception d'émotions simulées.....	10
2.4.2. La perception de l'état d'arrière-plan du corps.....	10
2.4.3. La simulation de la perception des émotions dans le cerveau	12
2.4.4. Comment se fait la perception des émotions ?.....	14
L'EXPRESSION DES ÉMOTIONS.....	15
3. DIFFÉRENTS TYPES D'EXPRESSION ÉMOTIONNELLE : DEUX APPROCHES	15
3.1. <i>La tripartition d'Ohala</i>	15
3.2. <i>Psychologie et linguistique : la rencontre de Scherer et Ladd</i>	16
3.2.1. Les différences d'approche.....	16
3.2.2. Une première expérience utilisant les deux approches	17
3.2.3. Résumé des résultats importants de l'expérience.....	18
3.2.4. Une deuxième expérience pour approfondir certains points	19
4. A LA RECHERCHE DES CARACTÉRISTIQUES ACOUSTIQUES ET PERCEPTIVES DE L'EXPRESSION VOCALE DE DIFFÉRENTES ÉMOTIONS	20
4.1. <i>Banse & Scherer 1996</i>	20
4.1.1. L'établissement du corpus.....	21
4.1.2. Le test de perception et ses résultats	21
4.1.3. L'analyse acoustique en vue de la reconnaissance automatique des émotions	23
4.2. <i>Leinonen et al. 1997</i>	24
5. LE RIRE ET LE SOURIRE	27
5.1. <i>Le sourire audible</i>	27
5.2. <i>Le sourire de Duchenne – marque d'amusement</i>	28
5.3. <i>Le rire</i>	29
5.3.1. Les caractéristiques acoustiques du rire	30
5.3.2. Le rôle social du rire.....	30
5.4. <i>Neurophysiologie du sourire et du rire</i>	31
5.4.1. Un centre déclencheur du rire : la SMA	31
5.4.2. Du sourire au rire – un continuum neurophysiologique	31
5.5. <i>En résumé</i>	31
QUELQUES PRINCIPES MÉTHODOLOGIQUES.....	33

6.	MÉTHODES D'ÉTABLISSEMENT D'UN CORPUS	33
6.1.	<i>La problématique : la contradiction d'un corpus contrôlé d'émotions naturelles</i>	33
6.2.	<i>Les différentes méthodes d'établissement des corpus</i>	33
6.2.1.	Émotions naturelles	33
6.2.2.	Émotions simulées par des acteurs	34
6.2.3.	Émotions suscitées en laboratoire	34
6.2.4.	Stimuli créés artificiellement par resynthèse	36
6.3.	<i>Présélection des énoncés</i>	36
7.	TEST DE PERCEPTION	37
7.1.	<i>Connaissances et meta-connaissances</i>	37
7.2.	<i>Quelques éléments de tests de perception en recherche sur la parole émotionnelle</i>	38
7.2.1.	Le questionnaire	38
a)	Type de questionnaire	38
b)	Choix fermé vs. choix ouvert	38
c)	Neutraliser les effets d'ordre	39
d)	Vérifier la fiabilité des juges	39
7.2.2.	Quelques éléments remarquables sur la présentation des stimuli	39
7.2.3.	L'exploitation des données	39
a)	Analyse statistique des données	39
b)	Trouver des réalisations type	40
	ANALYSE DU CORPUS " SOURIRE "	41
8.	L'EXPRESSION PROSODIQUE DE L'AMUSEMENT	41
8.1.	<i>Fondements théoriques de l'expérience</i>	41
8.1.1.	Amusement, énonciation souriante, et énonciation neutre	41
8.1.2.	Amusement spontané vs. simulé	42
8.1.3.	Amusement spontané vs. joué	42
8.1.4.	Sourire amusé vs. sourire social	43
8.1.5.	Hypothèses	43
8.2.	<i>Etablissement du corpus</i>	43
8.2.1.	La préparation des sessions d'enregistrement	44
8.2.2.	Les locuteurs	46
8.2.3.	Le protocole d'enregistrement	46
a)	L'enregistrement de l'amusement	46
b)	Productions " acteur ", " sourire social ", " sourire mécanique ", et " neutre "	47
c)	Simulation : doublage de l'enregistrement amusé spontané	47
8.2.4.	La numérisation des enregistrements	48
8.2.5.	Le corpus résultant	48
8.3.	<i>Test de perception</i>	51
8.3.1.	Planification du test de perception	51
8.3.2.	Un programme Supercard pour le test	52
8.3.3.	Le protocole du test de perception	53
8.3.4.	Commentaires des juges	53
8.4.	<i>Les résultats du test de perception</i>	54
8.4.1.	L'opposition amusé-neutre	54
8.4.2.	Les effets prosodiques de l'amusement	57
8.4.3.	Simulation par répétition	58
8.4.4.	Spontané et acteur	59
8.4.5.	Acteur d'amusement, de séduction, et sourire mécanique	60
	CONCLUSIONS ET PERSPECTIVES	64
	BIBLIOGRAPHIE	66
	INDEX	71

INTRODUCTION

Dans le domaine de la synthèse de la parole, un problème majeur est le manque de naturel de la parole produite. Tout en étant intelligible, elle manque d'expressivité. Beaucoup d'efforts ont été faits dans le domaine de la synthèse vocale sur l'intelligibilité : tout d'abord sur l'intelligibilité segmentale, et depuis plus récemment, mais avec intensité, sur l'intelligibilité suprasegmentale. Ainsi la prosodie est-elle souvent simulée dans les synthétiseurs dans ses fonctions de structuration linguistique. La dimension expressive de la prosodie naturelle a pour l'instant été peu étudiée (cf. les travaux sur la synthèse d'attitudes prosodiques de Morlec *et al.*, 1997). Cependant, de plus en plus de tentatives sont faites pour modéliser et implémenter des "caractéristiques du locuteur", notamment émotionnelles (Murray & Arnott, 1993 ; Scherer, 1996). Ces approches se servent de données sur les caractéristiques acoustiques de parole émotionnelle, pour façonner des prototypes de voix de synthèse.

Le but à long terme de cette étude sera l'implémentation de l'expression de l'amusement dans le système ICP audiovisuel de synthèse de la parole. Quelques tentatives existent déjà dans ce domaine en audio (Murray & Arnott, 1993) et vidéo (le système du KTH par exemple : Beskow, 1995). Dans la démarche anthropomorphique sous laquelle est envisagée la synthèse à l'ICP, il est apparu comme un préambule nécessaire de décrire par une analyse expérimentale une expression relativement facile à contrôler expérimentalement : celle de l'amusement ; et avant cela de retirer quelques hypothèses des travaux de plus en plus nombreux qui s'intéressent à l'expression vocale des émotions. Plusieurs problèmes apparaissent quand on survole rapidement les études sur l'expression vocale des émotions :

- une littérature fragmentée. Il existe nombre de petites études, de préférence opposant un petit nombre d'émotions "de base", mais, comme le souligne Scherer (1986), "*there has been neither continuity nor cumulativeness in the area of the vocal communication of emotion.*" (p.143) ;
- la complexité à dresser des théories de l'émotion (motivations éthologiques, origines, causes, fonctionnement physiologique, expression...). Souvent, l'étude de l'expression vocale émotionnelle se fait de manière plutôt empirique, sans vraiment se baser sur des hypothèses précises issues d'une théorie ;
- comment définir objectivement des états physiologiques (cognitifs ?) émotionnels sur lesquels porte la difficulté d'une définition sémantique par des étiquettes langagières ? Scherer (1986) déplore la grande difficulté à comparer les résultats de différentes études par manque d'informations sur ce que les auteurs entendent par une étiquette donnée. Peut-on se baser sur "l'expérience" des langues qui manipulent sous des formes variables les concepts d'émotions ? Par exemple, Scherer (1989) mentionne l'existence de listes de 550 adjectifs émotionnels pour l'anglais (Averill, 1975) et de 235 adjectifs pour l'allemand (Scherer, 1984a).

Ainsi, dans une première partie de ce travail, seront présentées brièvement quelques approches théoriques qui tentent de classer les émotions : Scherer (1989) développe un modèle des émotions rendant compte de leurs fonctions ; et Damasio (1994) distingue les émotions primaires (innées) des émotions secondaires (acquises au cours de l'histoire individuelle).

Ensuite nous recentrons cette revue sur quelques travaux théoriques et expérimentaux s'intéressant à l'expression des émotions, et tout particulièrement aux expressions de l'amusement du sourire au rire.

Nous examinerons ensuite des méthodologies utilisées dans la littérature pour l'établissement d'un corpus de parole émotionnelle, ainsi que quelques remarques de base pour l'établissement des tests de perception en parole émotionnelle.

La dernière partie concerne directement notre étude expérimentale de l'expression d'amusement. Nous avons d'abord récolté, en audiovisuel pour quatre locuteurs, un corpus d'expressions spontanées de l'amusement. Chaque énoncé "amusé" a été ensuite produit de différentes façons : en tant qu'acteur exprimant de l'amusement *vs.* une expression sociale (que nous avons identifié comme un sourire social de séduction) ; par répétition synchronisée *vs.* non synchronisée de l'expression spontanée ; avec un sourire mécanique (geste sans émotion causale) ; et enfin sans amusement. Les stimuli de ce corpus ont permis l'analyse perceptive des différents facteurs possibles d'opposition introduits dans le corpus.

Chapitre 1

LES EMOTIONS : FONCTIONS ET PHYSIOLOGIE — UN SURVOL BIBLIOGRAPHIQUE

1. Quelques notions théoriques sur les émotions

L'émotion est commune aux animaux et aux humains. Elles s'est en fait complexifiée au cours de l'évolution (Scherer, 1989). On peut l'aborder sous une approche phylogénétique¹ et comparative pour rendre compte de ses fonctions (Damasio, 1994 ; Scherer, 1989).

L'établissement d'une théorie convaincante des émotions pose problème depuis longtemps, étant donné la difficulté d'objectivation du phénomène. Ainsi, il existe un certain nombre d'approches différentes ; elles seront évoquées rapidement ci-dessous, avant une présentation plus approfondie de la théorie des processus composants de Scherer (1984b, 1989).

1.1. Les fonctions des émotions

Scherer (1989), se fondant sur une approche phylogénétique et comparative des émotions, décrit les émotions comme un mécanisme flexible d'adaptation à un environnement changeant. Contrairement aux réflexes, où l'association stimulus-réponse est immédiate, les émotions provoquent un découplage entre le stimulus et la réponse comportementale. Les émotions fonctionnent comme des agents intermédiaires entre l'environnement et le sujet. Les aspects les plus importants de ce processus sont, selon Scherer,

- “l'évaluation de la signification des stimulations ou événements du milieu par rapport aux besoins, projets ou préférences d'un organisme dans certaines situations (en particulier dans les processus d'apprentissage),
- la préparation physiologique et psychologique aux actions propres à répondre à ces stimulations du milieu et
- la communication des états et intentions de l'organisme à son environnement social ” (Scherer 1989, p. 101).

Les deux premières fonctions sont principalement physiologiques, centrées sur le sujet lui-même : elle lui permettent de mieux gérer la diversité possible de stimulations, en particulier par la focalisation de l'attention et la mise en place de priorités.

¹ La phylogénèse est la succession des espèces animales que l'on suppose descendre les unes des autres.

La troisième fonction des émotions, par contre, est une fonction de *communication* inhérente aux émotions. Ainsi, pour les espèces vivant en société, l'aspect visible des émotions rend possible tout un système de ritualisation, basé sur l'action probatoire :

“ l'expression motrice et l'ébauche d'action souvent lisible dans cette expression constituent des mouvements qui communiquent la réaction et l'intention d'action de l'individu. Par-là une sorte d'action probatoire est rendue possible : la réaction des autres fournit un feed-back qui permet une modification adaptative du projet d'action d'origine... certains modes d'expression semblent s'être spécialement constituées au cours de l'évolution en vue de la communication. ” (Scherer 1989, p. 102-103)

Un exemple flagrant pour l'action probatoire est la ritualisation du comportement agressif. Cette possibilité de transmettre des intentions de comportement à l'autre permet le développement de formes complexes d'organisation sociale. (L'aspect de la fonction communicative des émotions est approfondi par Ohala, 1996, voir 0).

Plus les stimulations à évaluer sont variées, plus le système émotionnel permettant leur traitement doit être complexe. Par conséquent, il y aurait chez l'homme une émotivité plus marquée que chez toutes les espèces animales. Ce paradoxe apparent est expliqué par Scherer par le fait que le contrôle social de l'affectivité fait obstacle à l'expression ouverte des émotions (les *display rules* d'Ekman & Friesen, 1969).

1.2. Un survol des différents types de théories

Scherer (1989), avant de présenter sa propre théorie des émotions, rappelle brièvement les théories existantes.

Les théories des émotions discrètes (p. ex. Izard, 1977b ; Tomkins, 1962) postulent un certain nombre d'émotions primaires (selon les auteurs entre sept et quinze environ). Les autres émotions seraient constituées par un mélange d'émotions primaires, comme, sur la palette d'un peintre, un nombre infini de couleurs peut être obtenu à partir de quelques couleurs fondamentales. Scherer (1989) objecte trois choses à une telle conception. D'abord, le choix des émotions primaires est relativement arbitraire, tant qu'il n'est pas justifié de manière objective, c'est-à-dire fondé sur des données biologiques ; ensuite, il est difficilement imaginable que même les émotions les plus opposées se mélangent. Enfin, personne n'a encore déterminé les proportions de ces mélanges pour la multitude d'états émotionnels.

La théorie de la cognition-activation (formulée par Schachter & Singer, 1962 ; ensuite défendue par Mandler, 1975) suppose qu'une activation physiologique non spécifique devient une émotion particulière en dépendance des cognitions existantes.

Les théories cognitives de l'évaluation (Arnold, 1960 ; Lazarus, 1966) considèrent également que la nature de l'émotion est déterminée par une évaluation cognitive. Le critère central d'évaluation est le caractère utile ou nocif pour l'organisme d'un stimulus.

L'approche dimensionnelle (Wundt, 1913) vise à décrire le vécu émotionnel au moyen de trois dimensions. Les dimensions envie-aversion (positif-négatif) et excitation-apaisement (actif-passif) sont retrouvées dans les recherches ultérieures à Wundt, tandis que sa troisième dimension, tension-soulagement, semble moins stable. La question est cependant plutôt où il

tension–soulagement, semble moins stable. La question est cependant plutôt où il faut situer ces dimensions : il n'est pas clair si le vécu émotionnel subjectif peut effectivement être décomposé selon ces dimensions, ou si ce sont simplement les mots, les dénominations verbales des émotions, qui se laissent positionner sur des dimensions sémantiques, ce qui reviendrait plus à une systématique linguistique que psychologique.

1.3. Le modèle des processus composants de Scherer

Le *component process model* de Scherer (1984b, 1989) décrit les émotions non pas en tant qu'états statiques, mais comme un ensemble de processus se déroulant dans des sous-systèmes de traitement. Dans ce modèle, les émotions correspondent à des changements d'état des sous-systèmes, ces changements étant déclenchés par des tests d'évaluation de la situation externe et interne (les *stimulus evaluation checks*, SECs). Les états des sous-systèmes peuvent être décrits en termes d'effet physiologique, et par la suite en termes de l'influence sur l'expression, en particulier sur l'expression vocale (Scherer, 1986, pour la description hypothétique ; Banse & Scherer, 1996, pour la vérification expérimentale des hypothèses).

Scherer (1989) distingue cinq sous-systèmes : un système d'assistance, influant sur les états neuro-endocrinien et végétatif de l'organisme ; un système d'action (état neuromusculaire) ; un système d'information (état de perception/souvenir etc.) ; un système régulateur (motivation actuelle, état de planification/décision), et enfin un système moniteur (état de conscience).

Les SECs sont au cœur de la théorie des processus composants. Il s'agit d'une série de tests pour vérifier en permanence la situation, et pour déclencher éventuellement des réactions bénéfiques pour l'individu. Les SECs sont supposés se dérouler toujours dans le même ordre temporel, du plus simple au plus élaboré. La totalité des SECs ne serait opérée que chez l'être humain adulte ; le développement des émotions, autant phylogénétique qu'ontogénétique, correspondrait à l'apparition de SECs de plus en plus complexes.

1. Test de nouveauté : le SEC le plus primitif est la vérification de la présence de changements. Ce test peut provoquer l'émotion de surprise, dans le sens d'un sursaut brusque à la suite d'un changement soudain important.
2. Test de valence hédonique : il s'agit ici d'une évaluation de bas niveau du caractère utile/nocif, agréable/désagréable d'un stimulus, entraînant des comportements d'approche (liés à l'émotion d'envie) ou d'évitement (liés à l'aversion). Les deux premiers SECs forment, suppose Scherer, des "*mécanismes précognitifs fondamentaux de traitement de la stimulation qui, sous des formes similaires, seraient communs à de nombreuses espèces animales supérieures*" (Scherer 1989, p. 122).
3. Test du rapport au but visé : ce SEC évalue le stimulus en termes de signification pour les projets de l'individu, s'il s'agit d'une aide ou d'un obstacle. Un stimulus intrinsèquement agréable (SEC 2) peut faire obstacle au but visé, déclenchant ainsi une émotion de frustration.
4. Test de la capacité de maîtrise : ce test évalue la capacité de l'individu à maîtriser une situation négative. Pour cela, il est indispensable de déterminer la cause d'une

stimulation. Si la situation semble surmontable sans danger pour l'organisme, l'émotion de colère est déclenchée ; sinon, l'émotion résultante est la peur, ou, lorsque la situation évaluée comme non maîtrisable est chronique, la dépression.

5. Test de la compatibilité aux normes et à l'image de soi : ce test, que Scherer suppose fonctionner exclusivement chez l'humain, confronte les stimulations aux normes sociales et à l'image de soi. Si le comportement de l'individu lui-même est évalué non conforme aux normes, il en résulte une émotion telle que la honte.

Les configurations typiques de résultats des SECs permettent de retrouver les émotions habituelles telles que colère, joie, tristesse etc. D'autre part, des configurations inhabituelles permettent de rendre compte de "mélanges" d'émotions : il s'agit de la co-présence d'évaluations à différents niveaux qui habituellement sont associées à des émotions différentes.

2. Neurophysiologie des émotions : Damasio 1994

2.1. Son approche générale

L'approche de Damasio se base sur une étroite interaction entre les pensées, les émotions, et le corps. Le titre de son livre, "L'erreur de Descartes – la raison des émotions" annonce clairement son propos : il associe le fonctionnement de la raison à celui des émotions. Ainsi, il se distingue de la perspective traditionnelle d'une dichotomie entre le néo-cortex qui abrite la pensée et les niveaux subcorticaux qui abritent les émotions, perspective qu'il caricature ainsi :

"Dit de façon la plus simple possible, les parties anciennes du cerveau, en bas, s'occupent de la régulation biologique fondamentale, tandis qu'en haut le néo-cortex réfléchit, avec sagesse et subtilité. Dans les étages supérieurs, au sein du néo-cortex, il y a la raison et la volonté, tandis qu'en bas, il y a les émotions et tout ce qui, banalement, concerne le corps." (traduction française (1995) de Damasio (1994), p. 170²)

Sa position à lui est la suivante :

"Les mécanismes neuraux sous-tendant la faculté de raisonnement, que l'on pensait traditionnellement situés au niveau *néo-cortical*, ne semblent pas fonctionner sans ceux qui sous-tendent la régulation biologique, que l'on pensait traditionnellement situés au niveau *subcortical*." (*ibid.*, p. 170)

2.2. Les émotions selon Damasio

Damasio (1994) oppose des émotions primaires – innées – à des émotions secondaires – acquises.

² Les numéros de page indiqués pour Damasio (1994) sont toujours celles de la traduction française (1995).

2.2.1. Les émotions primaires

Les émotions primaires sont pour lui toutes les réactions émotionnelles innées, pour l'homme ou pour les animaux. Le système limbique réagit à certains stimuli par une réponse émotionnelle comprenant des “*signaux vers les muscles du visage et des membres, [des] signaux vers le système nerveux autonome, [des] signaux vers les neurones modulateurs*” ainsi que des “*réponses endocrines et autres réponses chimiques par voie sanguine*” (p. 175). A l'intérieur du système limbique, ce sont l'amygdale et le cortex cingulaire antérieur qui jouent le rôle le plus important.

Notons que la réponse émotionnelle comporte, d'une part, différents signaux vers le corps proprement dit (signaux vers les systèmes moteur, nerveux autonome, et signaux chimiques par voie sanguine), d'autre part des signaux vers les neurones modulateurs, qui “*affectent considérablement le style et l'efficacité des processus cognitifs, et constituent une voie parallèle pour l'expression des émotions*” (p. 182).

Damasio accorde une utilité générale aux émotions primaires :

“... la réponse émotionnelle peut remplir quelques utiles fonctions : par exemple, elle peut permettre de se dissimuler rapidement à la vue d'un prédateur, ou de montrer à un concurrent que l'on est en colère.” (p. 175)

2.2.2. Les émotions secondaires

Du moins chez les êtres humains, il existe ensuite la possibilité de la “*perception de l'émotion*” (p.176), par quoi Damasio entend la possibilité de prendre conscience d'un lien entre la réaction émotionnelle et le stimulus qui l'a déclenché.

“... [les émotions secondaires] se manifestent à partir du moment où l'on commence à percevoir des émotions et à établir des *rapports systématiques entre, d'une part, certains types de phénomènes et de situations et, d'autre part, les émotions primaires.*” (p. 177-178)

Les émotions secondaires sont des réponses émotionnelles qu'on *apprend*, c'est-à-dire elles sont *acquises* au cours de “*l'histoire individuelle*” : à partir de la conscience du lien entre émotion primaire et stimulus déclenchant, on apprend “*la façon dont certaines situations ont généralement été couplées à certaines réponses émotionnelles au cours de l'histoire individuelle*” (p. 180).

Contrairement aux émotions primaires, ces émotions secondaires peuvent ensuite être déclenchées par l'imagination, c'est-à-dire par un acte purement mental. Si on *s'imagine* p. ex. la rencontre avec un ami qu'on n'a pas vu depuis longtemps, ou la mort inattendue d'un collègue, l'émotion secondaire correspondante est, selon Damasio, déclenchée avec les étapes suivantes :

1. la représentation consciente qu'on se fait d'une personne ou d'une situation, en partie verbale, en partie non-verbale ; les parties du cerveau impliquées sont “*divers cortex sensoriels fondamentaux*” (p. 180), et “*un grand nombre de cortex d'association de niveau élevé*” (p. 180) ;

2. la réponse du cortex préfrontal, automatique, involontaire et non consciente, selon les couplages entre situations et émotions, acquis au cours de l'histoire individuelle. C'est donc ici que les émotions secondaires sont déclenchées ;

3. non conscientes, automatiques et involontaires, les signaux en provenance du cortex préfrontal arrivent à l'amygdale et au cortex cingulaire antérieur, dans le système limbique. Ici sont déclenchées toutes sortes de signaux, comme pour les émotions primaires (voir 0), donc aussi bien des signaux vers le corps que des signaux influençant la manière dont se déroulent les processus mentaux. L'état émotionnel corporel résultant est signalé en retour aux systèmes limbique et somatosensoriel.

Les observations suivantes étayaient la différenciation hypothétique en émotions primaires et secondaires :

“ ... la perturbation des processus émotionnels chez les patients souffrant de lésions préfrontales concerne les émotions secondaires. Ces malades ne peuvent exprimer aucune émotion lorsqu'ils perçoivent les images évoquées par certaines catégories de situations et de stimuli, et par suite ne peuvent rien ressentir qui y corresponde. (...) Ces mêmes patients peuvent exprimer des émotions primaires, cependant (...) (ils montrent de l'effroi, si quelqu'un crie brusquement derrière eux, ou si leur maison tremble lors d'un séisme). Au contraire, des patients souffrant de lésions du système limbique, au niveau de l'amygdale ou du cortex cingulaire antérieur, montrent un déficit bien plus important, touchant à la fois les émotions primaires et secondaires, ... ” (p. 182)

Damasio résume que “ *l'émotion résulte de la combinaison de processus d'évaluation mentale, simples ou complexes, avec des réponses à ces processus, issues de représentations potentielles* ”. (p. 183) Cette position se rapproche en effet beaucoup du *component process model* de Scherer (1984b, 1989).

2.3. La spécificité des mécanismes neuraux sous-tendant les émotions : l'exemple du sourire

2.3.1. Emotions et l'hémisphère droit

Damasio cite toute une gamme d'articles³ confirmant que les structures neurales sous-tendant les émotions chez l'homme, ainsi que leur expression, sont localisées et se trouvent principalement dans l'hémisphère cérébral droit .

2.3.2. La différence entre le sourire spontané et le sourire volontaire

Un sourire spontané, à la suite d'une situation comique, n'a pas la même origine dans le cerveau qu'un sourire volontaire impliquant la commande motrice de contracter certains muscles du visage. Ceci a été mis en évidence à partir de patients souffrant de lésions cérébrales en différentes parties du cerveau :

³ Sperry, Gazzaniga, & Bogen (1969); Sperry, Zaidel, & Zaidel (1979); Gainotti (1972), Gardner, Browell, Wapner, & Michelow (1983); Heilman, Watson, & Bowers (1983); Borod (1992); Davidson (1992).

- lorsque le cortex moteur de l'hémisphère gauche du cerveau a été détruit, le côté droit du visage du patient est paralysé, ce qui résulte en une asymétrie de la bouche. Un sourire forcé, volontaire, augmente encore l'asymétrie ; cependant, un sourire involontaire, à la suite d'une remarque humoristique, est *normal, symétrique, et " ne différant en rien du sourire tel qu'il se manifestait chez cet individu avant la paralysie "* (p. 184) ;
- si le cortex cingulaire antérieur a été détruit au niveau de l'hémisphère gauche, on trouve le cas inverse : le sourire spontané est asymétrique, le sourire volontaire symétrique.

Par conséquent, *" la commande motrice des mouvements liés à l'émotion n'a pas la même origine que celle concernant les actes volontaires "* (p. 184).

Aussi pour les personnes saines, les différentes origines cérébrales des commandes ont pour effet une différence au niveau de l'expression pour le sourire spontané et le sourire volontaire. Comme G.-B. Duchenne l'a déjà observé en 1862, le sourire spontané met en jeu l'activation d'un muscle qu'il est impossible de contrôler volontairement :

" Duchenne avait montré que le sourire suscité par une joie réelle était réalisé par la contraction involontaire simultanée de deux muscles : le grand zygomatique [au niveau de la bouche] et l'orbiculaire palpébral inférieur [au niveau de l'œil]. Il a découvert en outre que ce dernier muscle ne pouvait être commandé que de façon involontaire ; il était impossible de le faire jouer volontairement. [...] En ce qui concerne le grand zygomatique, il peut être mis en jeu à la fois de façon involontaire et sous l'action de la volonté, et il est donc le moyen approprié pour réaliser des sourires de politesse. " (p. 186-187)

La distinction entre un sourire de Duchenne et d'autres sourires est étudiée en plus de détail par Ekman et al. (1990), voir 0.

2.4. La perception des émotions

Damasio s'intéresse à l'auto-perception des émotions, c'est-à-dire la perception qu'un individu a de sa propre émotion. Les informations sur l'état actuel du corps atteignent en continu le cerveau par une voie neurale et par une voie chimique (hormones et peptides dans le sang).

Damasio avance l'hypothèse suivante :

" C'est en ce processus de continuelle surveillance du corps, en cette perception de ce que votre corps est en train de faire tandis que se déroulent vos pensées, que consiste le fait de ressentir des émotions. [...] En d'autres termes, ressentir une émotion dépend de la juxtaposition d'une image du corps proprement dit avec une image de quelque chose d'autre, comme l'image visuelle d'un visage ou l'image auditive d'une mélodie. " (p. 189-190)

Il précise qu'il croit plutôt en une juxtaposition qu'en une fusion des deux images mentales, parce que *" cela permet de comprendre pourquoi il est possible de se sentir déprimé, alors que l'on pense à des personnes ou à des situations qui ne sont pas évocatrices de tristesse ou de dépossession, ou de se sentir gai, sans raison immédiate explicable. "* (p. 191) Le fait qu'il est possible de ressentir une émotion " inappropriée " à propos d'un objet peut donc être dû au fait qu'un facteur intervient, dans l'état du corps ou sa perception, qui n'est pas lié à l'objet en question, comme un changement purement physiologique.

Il ne faut pas oublier que les émotions influencent aussi *la manière même* dont on pense :

“ Fondamentalement, la tristesse ou la joie sont constituées par la perception de certains états corporels juxtaposés à certaines pensées (de quelque nature qu’elles soient), et d’une modification de la tonalité et de l’efficacité des processus de pensée. En général, étant donné que la tonalité des processus cognitifs et les signaux (positifs ou négatifs) relatifs à l’état du corps ont été déclenchés par le même système neural, ils tendent à être concordants [...]. Lorsque les signaux relatifs à l’état du corps sont de nature négative, la production des images mentales est ralentie, leur diversité est moindre, et le raisonnement est inefficace ; lorsque les signaux émanant du corps sont de nature positive, la production des images mentales est vive, leur diversité est grande, et le raisonnement peut être rapide, bien que pas nécessairement efficace. ” (p. 191)

2.4.1. *Tromper le cerveau – la perception d’émotions simulées*

La perception d’une émotion peut être engendrée par une action volontaire. Comme exemple, Damasio cite l’observation d’Ekman (1992) selon laquelle, à partir d’une expression du visage volontairement créée, on perçoit une émotion. Ekman (1992) a donné à des sujets normaux des instructions sur des mouvements des muscles faciaux à effectuer, qui “ *étaient susceptibles de leur conférer une expression typique d’une émotion, sans qu’ils le sachent.* ” (Damasio, p. 193) Les sujets déclaraient qu’ils éprouvaient effectivement l’émotion correspondante.

Ainsi, à partir d’une imitation grossière d’une expression émotionnelle, on percevrait l’émotion ? Damasio interprète les résultats d’Ekman comme suit :

“ Les expériences d’Ekman suggèrent soit qu’une partie de l’état corporel caractéristique d’une émotion suffit à permettre sa perception ; soit que cette partie suscite la reconstitution de l’état corporel complet, ce qui permet ensuite de ressentir l’émotion en question. ” (p. 193)

Néanmoins, il existerait une différence d’activité neuronale entre des expressions émotionnelles naturelles et volontaires :

“ De récents résultats d’électrophysiologie [Ekman & Davidson (1993)] montrent que les sourires commandés de façon volontaire ne sont pas accompagnés par les mêmes types d’ondes cérébrales que les sourires spontanés. ” (p. 193)

Cette différence au niveau des mesures cérébrales se retrouve aussi dans la conscience des sujets :

“ ... bien que disant ressentir l’émotion correspondant à l’expression faciale fragmentaire qu’ils se composaient, les sujets [d’Ekman (1992)] étaient parfaitement conscients que leur colère ou leur joie n’était pas due à telle ou telle circonstance. ” (p. 193)

2.4.2. *La perception de l’état d’arrière-plan du corps*

Il s’agit là d’une hypothèse de Damasio. La perception de l’état d’arrière-plan du corps serait la perception d’un état de fond, entre les épisodes émotionnelles, une sorte de cadre de référence à partir duquel nous expérimenterions des sensations plus particulières : une douleur, une émotion, etc.

“ La perception de l'état d'arrière-plan du corps est permanente, bien que vous la remarquiez à peine, puisqu'elle ne correspond à aucune partie spécifique du corps, mais plutôt à l'état global de la plupart de ses organes. ” (p. 198)

Il donne un exemple frappant de ce que serait une perception sans cette “ toile de fond ” :

“ on peut se demander ce qui arriverait (...) si, votre jambe vous faisant mal, et vous obligeant à la décroiser, vous n'avez perçu le malaise momentané au niveau de votre membre que de façon isolée dans votre esprit, au lieu qu'il ait été rattaché à un sens global du corps. ” (p. 198)

La perception de l'état d'arrière-plan du corps serait à la base de notre représentation du “ moi ” et servirait de “ *cadre de référence par rapport auquel nous pouvons prendre conscience des innombrables autres choses* ” (p. 201).

Quant à la localisation (ou plutôt non-localisation) de la représentation de l'état actuel du corps dans le cerveau, il remarque :

“ les représentations des états présents du corps figurent dans de multiples cortex somatosensoriels, (...) à la fois dans les hémisphères droit et gauche, (...) la représentation des états actuels du corps est formée en permanence par de nombreux éléments distribués sur un grand nombre de structures, à la fois corticales et subcorticales. ” (p. 196)

L'état d'arrière-plan du corps n'est jamais très positif ou très négatif, bien qu'il puisse être perçu comme plutôt plaisant ou déplaisant. S'il reste du même type pendant une certaine période, l'effet cumulé “ *contribue probablement à définir une humeur, bonne ou mauvaise ou indifférente* ” (p. 196)

Outre cette représentation de l'état actuel du corps “ *constituant des représentations " en prise directe ", continuellement changeantes, il existe des cartes plus stables de la structure générale du corps* ” (p. 197) Ces cartes plus stables donnent une idée de ce que le corps tend à être en général, “ *déconnecté* ” du moment présent.

Ces cartes plus stables semblent remplacer les cartes actuelles en cas d'absence d'information, comme p. ex. dans le cas d'un “ *membre fantôme* ” que certains patients sentent encore après une amputation, ou dans le cas encore plus extrême de l'anosognosie. Les patients anosognosiques présentent une paralysie du côté gauche du corps, due à une lésion particulière dans l'hémisphère droit, une lésion des aires somatosensorielles (Damasio, p. 94). Ils ne semblent ressentir aucune émotion, et ils peuvent perdre toute conscience de leur état.

“ Ils ne s'aperçoivent pas qu'ils sont paralysés (...). Ils ne peuvent se représenter les conséquences de leur état et ne sont pas préoccupés par leur avenir. Leur expression émotionnelle est minime, voire nulle, et leur capacité à ressentir les émotions est également nulle (...). Les lésions cérébrales chez ces patients anosognosiques ont pour effet d'interrompre les communications entre les régions qui sont le siège des cartes neurales relatives à l'état du corps ; souvent, elles détruisent aussi certaines de ces régions elles-mêmes. Ces dernières sont toutes situées dans l'hémisphère droit, bien qu'elles reçoivent des messages en provenance des côtés droit et gauche du corps. ” (p. 199)

Damasio présente une explication de l'anosognosie recourant à la notion de la perception de l'état d'arrière-plan du corps.

“ Ne pouvant disposer d’aucune information en provenance de leur organisme, les patients anosognosiques sont incapables de mettre à jour leur représentation du corps, et par la suite sont incapables de reconnaître, *via* leur système somatosensoriel de façon rapide et automatique, que leur paysage corporel a changé. Ils peuvent encore se représenter l’apparence qu’avait leur corps auparavant, une apparence qui n’est plus d’actualité. Et puisque celle-ci était satisfaisante, cela explique qu’ils la déclarent ainsi, en dépit de son état présent. ” (p. 199)

2.4.3. *La simulation de la perception des émotions dans le cerveau*

La perception des émotions implique donc normalement la perception de l’état actuel du corps, par voie neurale ainsi que par voie chimique. Apparemment, cette perception corporelle peut être simulée à l’intérieur du cerveau.

“ Il existe donc des mécanismes neuraux qui nous procurent des perceptions “ comme si ” elles provenaient d’états émotionnels, comme si le corps les exprimait véritablement. (...) je doute, cependant, que cette perception soit la même que celle émanant d’un état du corps réel. ” (p. 201-202)

Damasio soupçonne que ce mécanisme de simulation est acquis et qu’il se met en place par association répétée entre certaines situations et les réactions corporelles en ces situations. Le mécanisme de simulation est donc issu du mécanisme complet impliquant le corps.

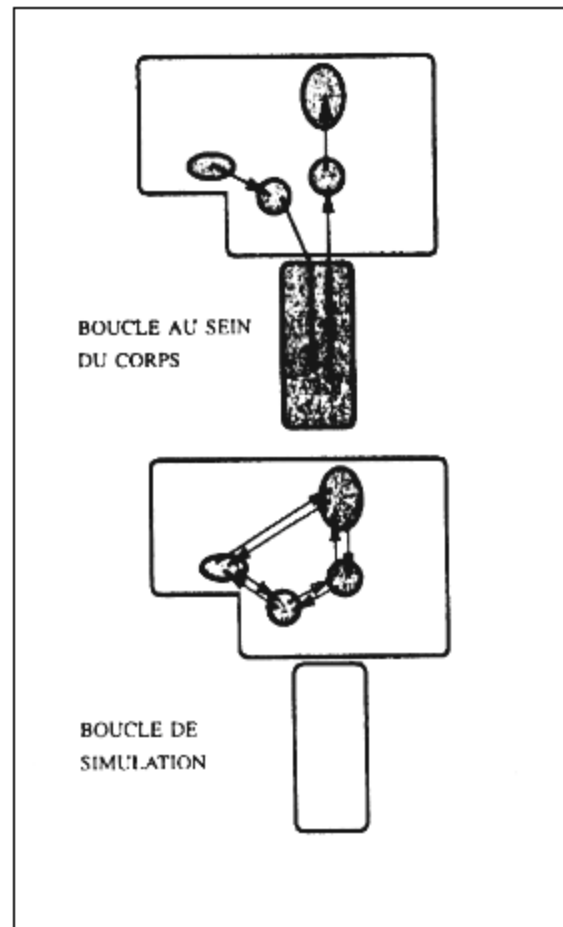


Figure 1. Schéma montrant les mécanismes de perception des émotions par le biais de processus se déroulant en “boucle au sein du corps” ou par le biais d’une “boucle de simulation”. Dans les deux images illustrant les deux sortes de “boucles”, le cerveau est représenté par la ligne noire fermée du haut, et le corps, par celle du bas. Les processus se déroulant dans la “boucle de simulation” court-circuitent complètement le corps. (selon Damasio, 1994, p. 202 de la traduction française de 1995)

Une conception concurrente avec celle de Damasio est la suivante :

“ le même agent cérébral qui a déclenché des changements au niveau du corps, envoie des signaux à un autre site cérébral (probablement le système somatosensoriel), l’informant des types de changements qui ont été demandés au corps. (...) les changements de l’état du corps se produiraient parallèlement à la perception des émotions, plutôt qu’ils n’en seraient la source. (...) la perception serait fournie dans tous les cas par le mécanisme de simulation. ” (Damasio, p. 204)

Damasio défend son point de vue par l’argument selon lequel l’émotion est induite par voie neurale *et chimique*, et qu’il est peu probable qu’un site cérébral informe un autre qu’il a induit une partie de la réaction par voie chimique. En outre, selon Damasio, le cerveau ne peut pas prédire précisément comment sera la réaction au niveau du corps ; par conséquent, sans prise en

compte du résultat réel au niveau du corps, on ne serait capable de ressentir que des émotions stéréotypés.

2.4.4. Comment se fait la perception des émotions ?

Les représentations neurales de l'état du corps ne sont pas la même chose que la perception de celui-ci ; il faudrait découvrir comment les représentations du corps sont intégrées au " moi ".

Pour ressentir une émotion à propos d'une situation, il faut associer la représentation de cette situation à la représentation de l'état du corps. Pour ceci, Damasio propose, outre ces deux représentations, une troisième composante, une représentation " tierce ", qui serait située dans les zones de convergence et qui, par des connexions bidirectionnelles avec les deux sources, servirait de lien entre les deux.

Chapitre 2

L'EXPRESSION DES EMOTIONS

Ce chapitre veut donner une introduction à plusieurs aspects théoriques importants, concernant les différents types d'expression d'émotions, la nature linguistique ou para-linguistique de cette expression, et les types d'informations véhiculées. Ensuite, le rire et le sourire sont examinés d'une manière plus détaillée.

3. Différents types d'expression émotionnelle : deux approches

3.1. La tripartition d'Ohala

Se basant sur des observations éthologiques⁴, Ohala (1996) a formulé l'hypothèse selon laquelle il faut distinguer trois types d'expression d'émotions⁵. Il propose de distinguer, d'abord, deux types d'expression émotionnelle proprement dite, que les humains partagent avec les animaux :

- “signals which do reflect the (psycho)physiological state of the signaler” (Ohala, 1996).

Ce type d'expression reflète un état intérieur émotionnel et, par cette fonction, semble correspondre au “symptôme” de Bühler (1934).; celui-ci est une adaptation à une situation donnée. L'adaptation est importante pour la survie : p. ex. la peur est une activation physiologique en cas de danger, préparant à la fuite. L'expression, quant à elle, est vu comme un produit secondaire inévitable et non intentionné de cette adaptation : l'activation de tous les muscles fait entre autres trembler les mains et la voix.

- “signals that convey messages where the transmission of the message itself has survival value” (Ohala, 1996).

Ce type d'expression émotionnelle n'a pas son origine dans un état intérieur, mais a pour but d'influencer le destinataire. La fonction s'apparente donc à celle du “signal” de Bühler. Tout message du type menace/non-menace appartient à ce type, comme le sourire, qui suggère que l'émetteur n'est pas dangereux, ou la voix basse et les sourcils baissés signalant que l'émetteur est grand et, par conséquent, une menace pour un adversaire potentiel. Ces expressions devraient, suppose Ohala, être universelles dans différentes cultures et peut-être dans différentes espèces.

⁴ Etude du comportement des espèces animales dans leur milieu.

⁵ Ces trois types, dans le domaine de l'expression émotionnelle, semblent correspondre aux trois fonctions du signe langagier du modèle fonctionnel de K. Bühler (1934). Selon celui-ci, le signe langagier peut être : “symptôme”, pour l'expression de l'état intérieur du locuteur ; “signal”, pour faire appel à l'interlocuteur ; ou “symbole”, pour désigner un objet ou un fait de la réalité.

L'hypothèse du “code fréquentiel” d'Ohala (1983, 1984, 1994) appartient au domaine du “signal”. Selon cette hypothèse, le F0 sera utilisé pour signaler une agression, par une voix basse, suggérant un émetteur grand et par conséquent dangereux, ou au contraire pour signaler une soumission, par un F0 haut, associé normalement à un individu petit et donc moins fort et moins dangereux. L'idée semble exemplaire pour le fonctionnement du “signal” : une propriété comme le F0 varie à l'origine avec la taille de l'émetteur, de manière naturelle et automatique, et la taille de l'émetteur détermine ses chances de gagner une lutte. Dans un deuxième temps, cette propriété est, dans la mesure des capacités de l'émetteur, découplée de sa cause originelle et utilisée comme signal, avec *l'intention de faire croire* le récepteur que l'émetteur a la propriété qu'indique le signal émis.

Enfin, il y a un troisième type d'expression émotionnelle dans un sens plus large : l'expression attitudinale, propre aux humains :

- “ways that humans convey their attitudes about the receiver, about the content or referent of their utterance, or about themselves” (Ohala, 1996).

La fonction de cette expression attitudinale étant de donner une information *à propos* d'un objet quelconque, elle semble correspondre au “symbole” de Bühler. Les attitudes incluent l'ironie, le sarcasme, l'indifférence etc. Contrairement aux deux premiers types, l'expression attitudinale ne semble pas issue d'un comportement bénéfique à la survie. Ohala suppose que ce type d'expression est acquis et qu'il varie par conséquent d'une culture à l'autre, voire d'un individu à l'autre. Et, très important pour le sujet qui nous intéresse ici, seulement ce troisième type d'expression émotionnelle aurait besoin “d'une grande quantité de contexte pragmatique et linguistique de haut niveau” pour être communiqué adéquatement.

3.2. Psychologie et linguistique : la rencontre de Scherer et Ladd

La rencontre d'un psychologue expérimental, K. R. Scherer, et d'un linguiste, D. R. Ladd, (Scherer et al., 1984 ; Ladd et al., 1985), a confronté deux approches différentes d'un même domaine, celui de l'étude des aspects non-verbaux dans la communication parlée. Les deux approches font des suppositions de base différentes quant à la nature de l'expression vocale des émotions. Par la suite, les méthodes utilisées ne seront pas les mêmes, ni les résultats obtenus. Ce fait est une bonne illustration du principe selon lequel la recherche scientifique n'est jamais neutre et ne permettra jamais de trouver *la vérité* : les résultats qu'on obtient dépendent des a priori sur lesquels on se base.

3.2.1. Les différences d'approche

Les études expérimentales qui cherchent à isoler des paramètres acoustiques en “contrôlant” le texte verbal des énoncés supposent un canal vocal parallèle aux paroles, par lequel l'information sur les émotions s'exprime indépendamment du contenu verbal. Les mesures habituellement effectuées dans des études statistiques (moyenne de F0, registre, ...) sont conformes à un modèle statistique où les jugements des auditeurs sont basés sur une *covariance de variables continues* : s'il est possible d'identifier des indicateurs acoustiques pour l'expression d'une émotion, on

suppose que plus l'émotion est forte, plus l'indicateur sera présent. Ceci mène à faire juger, dans des tests de perception, la force émotionnelle d'un stimulus sur une échelle de graduation.

Par contre, beaucoup de descriptions linguistiques affirment que l'intonation fait passer des significations émotionnelles uniquement en interaction avec certaines propriétés linguistiques verbales, ce qui signifie que des paramètres intonatifs identiques peuvent s'interpréter très différemment selon le texte avec lequel elles sont utilisées. Les linguistes partent du principe que l'intonation implique un certain nombre de distinctions catégorielles. En termes statistiques, les jugements des auditeurs sont par conséquent basés sur des *configurations de variables catégorielles*. Les questionnaires utilisés visent donc une décision catégorielle.

3.2.2. Une première expérience utilisant les deux approches

Pour tester ces deux approches de base, Scherer et al. (1984) ont mené des tests de perception sous différentes conditions avec un corpus de questions produits par 11 locuteurs allemands en contexte naturel. La "force affective" d'un énoncé se mesure en psychologie en termes d'impressions des auditeurs, d'où la nécessité de passer par des tests de perception. Le questionnaire permettait un choix ouvert parmi 9 adjectifs, choisis de la manière suivante. D'une liste d'environ 250 adjectifs allemands, cinq juges connaissant le corpus ont choisi les termes applicables au corpus, et les expérimentateurs ont choisi neuf de ces termes dont les sens se recoupaient aussi peu que possible.

Une première expérience visait à déterminer les rôles du texte et des indices non-verbaux pour les choix des juges. 32 auditeurs jugeait les stimuli audio, 24 les transcriptions écrits. Les résultats montrent que les indicateurs non-verbaux sont essentiels pour la communication des émotions : pour un seul adjectif, il existait une corrélation entre les jugements en condition audio et en condition transcription, ce qui indique que pour cet adjectif, des indicateurs verbaux étaient présents.

Dans un deuxième temps, le contenu verbal était masqué par différentes méthodes : filtrage à la valeur maximale de F0 pour chaque énoncé, pour laisser passer F0 seulement ; *random splicing* (découper le signal en morceaux et recoller au hasard) ; et parole inversée. Ces trois versions dégradées des stimuli, ainsi qu'une version audio pleine, étaient présentés à 18 sujets. Pour tous les adjectifs, les jugements des versions pleines étaient corrélées avec les jugements d'au moins une des versions préservant le timbre (*random splicing* et parole inversée). Ceci suggère fortement que le timbre véhicule beaucoup d'information émotionnelle, indépendamment du contenu verbal et malgré les distorsions importantes des contours de F0 et de l'énergie. La version filtrée, retenant essentiellement le contour de F0, était corrélée avec la version pleine pour un seul adjectif, POLI. Une interprétation selon le modèle *covariance* doit conclure que le contour de F0 ne contribue que très peu au message émotionnel de l'énoncé.

C'est dans la troisième phase que cette vision de F0 est révisée. Les résultats des deux premières parties ont été analysés de nouveau, d'une manière différente, pour tester l'approche configurationnelle :

“ [The aim was to show] that given appropriate hypotheses about the categorical organization of intonation, F0 cues can indeed be shown to play a significant role in conveying affect ” (Scherer et al., 1984, p. 1351).

Les contours intonatifs ont donc été classifiés dans les catégories MONTANT et DESCENDANT, selon le mouvement final de F0. Les énoncés verbaux ont été classés en questions partielles (*wh-questions*) et questions totales (*yes/no questions*). Une analyse statistique avec les deux types de phrase, les deux types de contour de F0, ainsi que les paramètres moyenne de F0 et registre (*F0 range*) a abouti aux résultats suivants :

- les choix “ impatient ” et “ détendu ” ont été faits sur la base des paramètres continus MOYENNE DE F0 et REGISTRE ;
- pour les choix “ reproche ” et “ agressif ”, la configuration de TYPE DE QUESTION et TYPE DE CONTOUR était important, c'est-à-dire seulement la combinaison d'une question totale avec un contour descendant aboutissait à ces jugements. Ceci est donc conforme aux prédictions du modèle configuration ;
- Pour un troisième type de jugements, “ gentil ”, “ plein de compréhension ”, et “ poli ”, les jugements étaient influencés en même temps par la MOYENNE DE F0 et l'interaction entre TYPE DE QUESTION et TYPE DE CONTOUR.

Ces résultats confirment la validité du point de vue configurationnel, selon lequel la configuration catégorielle de F0, en interaction avec le contenu verbal, contribue au message émotionnel. En même temps, il semble que les différents types de paramètres (continus et catégoriels) ne participent pas au même point pour les différents jugements : “ *the continuous variables appear to reflect states of the speaker related to physiological arousal, while the more linguistic variables tend to signal speaker attitudes with a greater cognitive or attitudinal component, such as friendliness or reproach* ” (Ladd et al., 1985, p. 435).

3.2.3. Résumé des résultats importants de l'expérience

Résumons les résultats principaux obtenus par Scherer et al. (1984). D'une part, les deux modèles, covariance et configuration, arrivent à identifier des paramètres contribuant à l'identification d'un message émotionnel ou attitudinal. Le modèle de covariance, supposant des paramètres continus, indépendants du contenu verbal, semble décrire adéquatement les effets du timbre, notamment. Le modèle de configuration, étant donné des hypothèses adéquates sur la catégorisation, peut décrire l'interaction entre le contenu verbal et le type de contour intonatif. Il est cependant important de noter que pour F0, il existe des paramètres catégoriels (type de contour) et des paramètres continus (moyenne de F0, registre). Ces derniers semblent contribuer au message émotionnel seulement en présence du contenu verbal.

Les variables continues semblent refléter des états du locuteur lié à l'excitation physiologique, pendant que les variables catégorielles, plus linguistiques, seraient impliquées dans le signalement des attitudes du locuteur.

3.2.4. Une deuxième expérience pour approfondir certains points

Dans un article suivant (Ladd et al., 1985), les auteurs reprennent de manière bien contrôlée les points clé de la première expérience. Ils se basent sur les hypothèses suivantes.

1. Le type de CONTOUR intonatif, le REGISTRE (*F0 range*) et le TIMBRE ont des effets indépendants sur des jugements émotionnels.
2. Le REGISTRE et le TIMBRE reflètent des états d'excitation (*arousal*), pendant que des différences de type de CONTOUR signalent des différences d'attitude "cognitive".
3. Le REGISTRE fonctionne comme une variable continue, de sorte à ce que des changements dans le REGISTRE sont directement corrélés avec des changements dans l'intensité des jugements émotionnels.

Trois petites expériences ont été menées pour vérifier ces hypothèses.

Dans un premier temps, un seul locuteur a prononcé trois phrases en allemand, avec deux timbres (normal, strident)⁶. Par resynthèse de la courbe mélodique⁷, chacun de ces 6 énoncés était préparé avec deux CONTOURS intonatifs et deux REGISTRES, menant à un total de 24 stimuli. Ceux-ci ont été présentés dans un test de perception à 23 sujets allemands, dans deux sessions, avec deux questionnaires différents. Le premier questionnaire consistait de 5 échelles bipolaires pour les états liés à l'excitation (détendu/excité, etc.), le deuxième de 5 échelles unipolaires pour les attitudes cognitives (plus ou moins de reproche, etc.). Les résultats étaient, d'une part, une confirmation de l'hypothèse 1 : il n'y avait pas d'interaction entre CONTOUR, REGISTRE, et TIMBRE dans leurs effets sur les jugements. D'autre part, les résultats relativisent l'hypothèse 2. S'il est vrai que les effets les plus forts de REGISTRE et de TIMBRE se trouvaient pour les jugements d'excitation, il y en avait cependant aussi pour les attitudes cognitives. Quant aux effets de CONTOUR, ils n'étaient significatifs que pour trois des cinq attitudes, et il existait un effet significatif même plus fort sur trois des cinq jugements d'excitation. Il ne semble donc pas possible de distinguer aussi strictement les effets des paramètres selon leur fonction.

Une deuxième expérience a été construite pour généraliser les résultats obtenus, pour plusieurs locuteurs. Trois locuteurs ont lu les trois phrases de l'expérience 1, mais sans faire une distinction de timbre. Comme avant, deux registres et deux contours mélodiques ont été construits sur chaque énoncé par resynthèse. Contrairement à l'expérience 1, un seul questionnaire était utilisé pour le test de perception, contenant 8 échelles unipolaires, quatre pour les états d'excitation, quatre pour les attitudes. Les 17 juges allemands jugeaient les stimuli en une seule session. Les résultats indiquent un premier effet principal pour le LOCUTEUR et pour la variable TEXTE (= les trois phrases). Cependant, ces variables ne sont pas en interaction avec les variables acoustiques, ce qui indique que tout effet des variables acoustiques qui serait répliqué de l'expérience 1 peut être traité comme un effet relativement général. C'est le cas pour le REGISTRE. Les résultats de la première expérience sont reproduits : l'effet le plus grand est celui

⁶ L'impossibilité de modifier artificiellement le timbre, liée à l'énorme difficulté de seulement le mesurer, force les auteurs à faire produire les différents timbres par le locuteur, ce qui rend la description du timbre relativement vague.

⁷ La resynthèse du contour mélodique était à l'époque une méthode très récente, utilisée entre autres par Liberman & Pierrehumbert (1984). Pour plus de détails sur la resynthèse, voir aussi 0 (0).

sur le jugement d'excitation, et les effets se répartissent à travers les états d'excitation et les attitudes. Par contre, pour CONTOUR, un seul effet a été trouvé, pour le jugement d'emphase. De nouveau, il n'y avait pas d'interaction significative entre les variables acoustiques, ce qui soutient l'idée de leur indépendance.

La troisième expérience visait la question du fonctionnement continu ou catégoriel du REGISTRE. Deux locuteurs prononçaient deux phrases, qui par resynthèse étaient transformées en un continuum de cinq niveaux de REGISTRE. Le questionnaire et le déroulement du test de perception étaient identiques à l'expérience 2, avec 25 juges. De nouveau, l'effet important du REGISTRE sur tous les jugements a été répliqué ; comme avant, l'effet le plus fort était celui pour le jugement d'excitation. La visualisation des moyennes des jugements pour les cinq niveaux de REGISTRE, ainsi qu'une analyse statistique, suggèrent fortement une perception continue des changements de REGISTRE. Cependant, pour l'un des deux locuteurs, le jugement moyen décroît du quatrième au cinquième niveau de REGISTRE, ce qui pourrait indiquer un changement de perception pour des valeurs extrêmes de REGISTRE.

En résumé, cette expérience de Ladd et al. (1985) apporte les conclusions suivantes. Premièrement, les effets de CONTOUR, REGISTRE et TIMBRE sur les jugements émotionnels des auditeurs sont indépendants les uns des autres. L'absence d'interactions entre ces variables indique un fonctionnement essentiellement additif, c'est-à-dire, dans la terminologie de Scherer et al. (1984), selon des "canaux parallèles". Deuxièmement, le TIMBRE et le REGISTRE ont un effet sur la plupart des jugements, l'effet le plus fort étant celui sur le jugement d'excitation. Cet effet a été reproduit pour le REGISTRE dans toutes les trois expériences. Cet effet plus fort sur les jugements d'excitation semble être le seul résultat soutenant l'hypothèse 2 ; une distinction claire entre des variables spécifiques pour les états d'excitation et pour les attitudes cognitives n'a pas pu être trouvée. Troisièmement, le REGISTRE semble avoir un effet perceptif continu sur tous les jugements, états d'excitation et attitudes confondues. Enfin, comme il n'existe pas d'interaction entre les variables acoustiques REGISTRE et CONTOUR d'une part et les variables LOCUTEUR et TEXTE d'autre part, ces résultats peuvent probablement être généralisés pour différents locuteurs et différents textes.

4. A la recherche des caractéristiques acoustiques et perceptives de l'expression vocale de différentes émotions

4.1. Banse & Scherer 1996

Scherer (1986) se base sur le *component process model* des émotions (voir 0) pour prédire, pour différentes émotions, les changements physiologiques déclenchés et les effets sur la vocalisation. Cela lui permet de prédire les propriétés acoustiques d'une douzaine d'émotions.

L'expérience de Banse & Scherer (1996) a comme but principal la vérification des prédictions de Scherer (1986) au sujet des profils acoustiques des expressions vocales d'émotions. En outre, elle vise

- à démontrer qu'il existe plusieurs dimensions de similitude perceptive entre émotions, et non pas une seule, l'excitation physiologique ;
- à trouver des indices pour l'existence ou non de prototypes mentaux qui seraient utilisés par des auditeurs pour reconnaître l'expression vocale d'une émotion ;
- à classer avec des méthodes statistiques les stimuli de parole émotionnelle, en se basant sur les profils acoustiques obtenus, et à comparer les résultats avec ceux des auditeurs humains.

4.1.1. L'établissement du corpus

14 émotions ont été choisies : les quatre paires d'émotions *hot anger/cold anger*, *elation/happiness*, *desperation/sadness*, et *panic fear/anxiety* représentent les quatre "familles" d'émotions largement étudiées (colère, joie, tristesse, peur), à deux degrés d'intensité ; ensuite, un nombre d'émotions n'appartenant pas à ces "familles", *interest*, *boredom*, *shame*, *pride*, *disgust*, et *contempt*. Comme support vocal, deux phrases ont été construites qui sont composées de phonèmes de plusieurs langues indo-européennes, mais qui n'ont aucun sens. Deux scénarios de mise en situation pour chaque émotion ont été choisis dans un corpus de situations antécédentes à des émotions, issu de plusieurs grandes études interculturelles. Douze acteurs professionnels allemands, six hommes et six femmes, ont produit les émotions, en étant enregistrés en audiovisuel⁸.

Pour assurer la qualité des énoncés utilisés dans le test de perception et l'analyse acoustique, une présélection des énoncés a été faite par des experts, des acteurs en formation. Les 1344 enregistrements ont été présentés dans les conditions audio seul, vidéo seule, et audiovisuel, triés par émotion exprimée. Les experts les jugeaient selon authenticité et reconnaissabilité, avec des notes scolaires⁹. Seuls les énoncés ayant obtenu une bonne note ont été retenus, ce qui a réduit le nombre d'énoncés à 280. Ensuite, pour des raisons formelles d'homogénéité de l'expérience, les expérimentateurs ont réduit le nombre d'énoncés au nombre final de 224. Ce processus de présélection amène déjà un premier résultat : la capacité d'exprimer des émotions de manière convaincante n'est pas la même pour tous les acteurs. Parmi les six acteurs masculins, trois ont fourni 88 % des stimuli. Parmi les actrices, la distribution était moins déséquilibrée.

4.1.2. Le test de perception et ses résultats

Dans un test d'identification perceptive, les 280 stimuli de la présélection étaient présentés à douze étudiants. Le test était divisé en deux sessions à cause du grand nombre de stimuli.

⁸ Dans deux études parallèles, l'expression faciale et corporelle des émotions est étudiée : Ellgring (1995) ; Wallbott (1995).

⁹ Les notes du système scolaire allemand vont de 1 à 6. La meilleure note est le 1 ; seulement 5 et 6 sont considérés comme au-dessous de la moyenne.

Chaque stimulus, audio seul, était présenté une seule fois, et les juges devaient l'identifier parmi les 14 dénominations émotionnelles. L'intérêt de ce test de perception n'était pas seulement de mesurer le taux de reconnaissance absolu (moyenne à travers les émotions = 48 %) et les différences de taux de reconnaissance pour les différentes émotions (de 78 % pour *hot anger* à 15 % pour *disgust*) ; pour le but que s'étaient donné les auteurs, il était aussi très important d'analyser les erreurs afin de déterminer les dimensions de similitude perceptive. Les résultats montrent que ce n'est pas uniquement le degré d'ACTIVATION PHYSIOLOGIQUE qui caractérise les émotions dans l'expression vocale, comme le soutiennent certains (p. ex. Pakosz, 1982, selon Alt, 1997). La matrice de confusion¹⁰ de Banse & Scherer, dont nous avons visualisé les principales tendances dans la Figure 2, montre que ce sont trois dimensions, ACTIVATION, QUALITE (type d'émotion), et VALENCE (positif vs. négatif) qui sont récupérées perceptivement dans l'expression vocale.

D'une part, cette présentation graphique permet de voir les trois dimensions de similitude :

- la dimension de L'ACTIVATION se voit par une confusion considérable entre les émotions les plus intenses (*hot anger, elation, desperation, panic fear*),
- la dimension de QUALITE par les confusions entre les paires *hot anger / cold anger, desperation / sadness, et panic fear / anxiety* (les émotions *elation* et *happiness* n'ont cependant pas été confondues – elles n'ont apparemment pas été considérées comme une “ famille d'émotions ” partageant la même qualité),
- et la dimension de VALENCE se voit par le fait que les émotions positives (*happiness, pride, interest*) n'appartenant pas à une même famille sont beaucoup confondues entre elles, de même les émotions négatives entre elles (*sadness et boredom, anxiety et desperation, contempt et cold/hot anger*), et que les émotions positives et négatives ne sont que peu confondues.

D'autre part, le fait que certaines confusions ne sont pas symétriques plaide pour l'existence de *prototypes acoustiques* pour certaines émotions et non pas pour d'autres. Aussi la “ force d'attraction perceptive ”, représentée par la taille différente des points dans la Figure 2, indique une préférence pour certains choix, qui pourrait être due à l'existence de *prototypes acoustiques* pour ces émotions. Par exemple, *elation* a souvent été reconnu comme *hot anger*, mais jamais *hot anger* n'a été reconnu comme *elation*. Banse & Scherer supposent une analyse descendante (*top-down*) pour les émotions à prototype, comme *hot anger*, facilitant la reconnaissance, et une analyse ascendante (*bottom-up*) pour les émotions sans prototype acoustique, comme *elation*, menant plus facilement à de mauvaises interprétations. Ce serait donc pour leur profile acoustique typique que des émotions à prototype fonctionneraient comme attracteurs perceptifs.

¹⁰ Une matrice de confusion rend compte des erreurs dans un test d'identification. Il s'agit d'un tableau dans lequel les réponses attendues (les “ bonnes ” réponses) figurent en tête des colonnes, les réponses effectivement données en tête des lignes (ou vice versa). Les valeurs dans le tableau donnent, pour chaque réponse attendue, les effectifs ou les pourcentages des réponses effectivement données. Par conséquent, si tous les juges avaient toujours donné la bonne réponse, les valeurs le long de la diagonale seraient de 100%, et tous les autres de 0. La diagonale indique donc le taux de bonnes réponses, tandis que les autres valeurs permettent de voir quels types de stimuli ont souvent été confondus, indiquant ainsi une similitude perceptive.

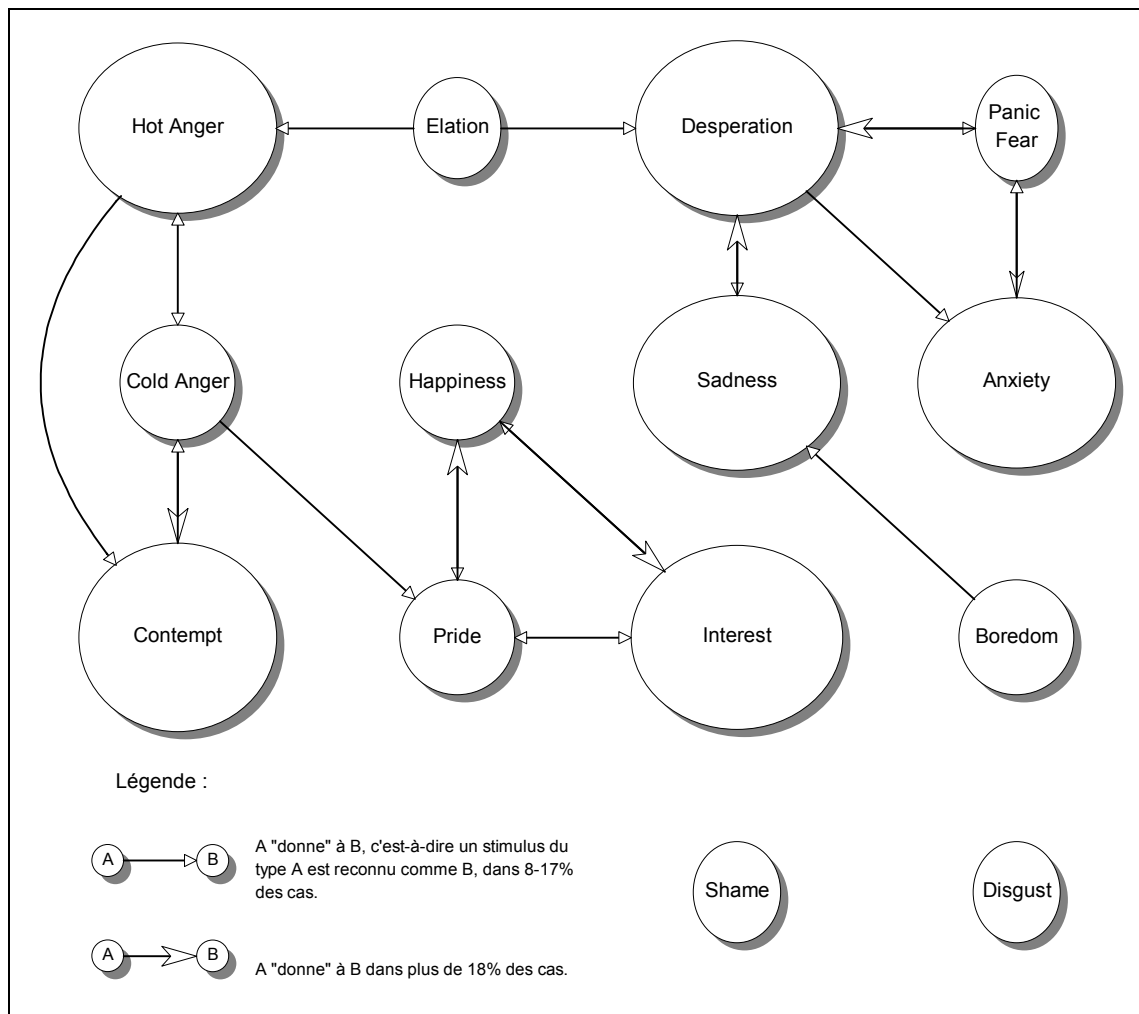


Figure 2. Une présentation graphique de la matrice de confusion de Banse & Scherer (1996).

Les points représentent les 14 catégories du test de perception. La taille d'un point représente le nombre de fois où la catégorie associée a été choisie, indépendamment de la justesse des choix. Ce nombre se calcule dans la matrice de confusion comme la somme des pourcentages dans chaque ligne "émotion reconnue". Cette valeur est un indicateur de l'attraction de la catégorie dans les choix des locuteurs. Un point de taille moyenne indique une valeur entre 80 et 120, un petit point une valeur au-dessous de 80, et un grand point une valeur au-dessus de 120. Les flèches entre points représentent des confusions entre catégories. Une flèche d'une émotion A vers une autre émotion B signifie qu'un stimulus du type A est reconnu comme B dans au moins 8% des cas, donc plus souvent que par hasard (7%). Les confusions multiples des catégories moins bien reconnues, *shame* et *disgust*, ne sont pas représentées. Les quatre familles d'émotions, colère, joie, tristesse, et peur, se trouvent dans les deux premières lignes, avec l'émotion plus intense de chaque famille en haut. La troisième ligne contient des émotions n'appartenant pas à ces familles. Il est remarquable qu'il est possible de représenter les confusions ainsi, d'une manière bidimensionnelle, étant donné qu'a priori, chaque émotion aurait pu être confondue avec chaque autre.

4.1.3. L'analyse acoustique en vue de la reconnaissance automatique des émotions

Dans une analyse acoustique, 29 paramètres acoustiques ont été extraits pour chacun des 224 énoncés sélectionnés. Les paramètres étaient calculés par moyennage à travers l'énoncé : moyenne, écart-type, quartiles de F0, moyenne de l'énergie, durée totale de segments voisés et

de pauses, ainsi qu'un grand nombre de valeurs pour le spectre acoustique "moyen" pendant l'énoncé, calculées séparément pour les périodes voisées et non-voisées¹¹.

Ces paramètres donnent une caractérisation acoustique des 14 émotions étudiées¹². Ces valeurs sont ensuite comparées aux prédictions de Scherer (1986). Tandis que certains résultats étaient comme prédit, d'autres différaient, permettant ainsi de préciser le modèle : "... *the empirical findings obtained in this study allow one to begin to disambiguate cases in which opposing physiological mechanisms rendered clear predictions impossible. This empirical input into the theoretical model is an essential part of the interaction between empirical data gathering and theory building that is often mentioned but rarely practiced in this area.*" (Banse & Scherer, 1996, p. 631)

Pour vérifier si les paramètres acoustiques obtenus permettraient une classification automatique des émotions, les stimuli ont été classifiés avec deux méthodes statistiques différentes (*jack-knifing* et *discriminant analysis*). Les résultats ont été comparés avec ceux des auditeurs humains. Les taux moyens de reconnaissance sont de 40 % environ pour chacune des méthodes statistiques, vs. 48 % pour les humains. Plus important, aussi les taux de reconnaissance différentiels sont comparables pour beaucoup d'émotions, c'est-à-dire que des émotions bien reconnues par les auditeurs humains étaient aussi bien reconnues par les méthodes statistiques, sauf pour trois exceptions. Même les matrices de confusions entre émotions semblent similaires ; cependant, ce ne sont pas les mêmes stimuli qui sont classifiés dans les mêmes cases dans les trois ensembles de données. Les méthodes statistiques de classification ne semblent donc pas vraiment répliquer le processus d'inférence des auditeurs humains.

4.2. Leinonen et al. 1997

Leinonen et al. (1997) ont étudié l'expression de dix connotations émotionnelles/motivationnelles dans un énoncé constitué d'un seul mot. Le but de l'expérience était d'étudier des vocalisations humaines de longueur comparable à des vocalisations de singes. Des expériences antérieures suggéraient que l'homme et les singes partagent des indices acoustiques pour peur, agression, dominance, soumission, satisfaction, et neutralité émotionnelle. Pour tester cette hypothèse, des vocalisations courtes étaient requises. Le nom d'une personne a été jugé approprié, parce que des locuteurs expriment leur état émotionnel/motivationnel souvent en énonçant le nom d'une personne, et parce que un nom est compatible avec beaucoup de connotations différentes.

16 locuteurs ont prononcé le nom [saara] dans chacune des connotations suivantes : *naming, commanding, angry, frightened, pleading, astonished, content, admiring, scornful, et sad*. Ce

¹¹ Ces moyennes de spectres à travers une phrase entière, même en séparant périodes voisées et non-voisées, semble difficilement admissible du point de vue d'un phonéticien. Quelle est la moyenne d'une occlusive voisée et d'une voyelle ? Il faut donc envisager ces mesures relatives au but que les auteurs se sont fixées, c'est-à-dire de mesurer les *différences* de timbre entre les 14 émotions, sur deux phrases standard. Même si donc le spectre moyen lui-même à travers toute la phrase n'a pas de sens, les différences de spectre moyen entre émotions peuvent donner des indications sur les différences de timbre.

¹² Dans un autre article exploitant ces mêmes données différemment, Johnstone, Banse & Scherer (1995) séparent, pour chaque émotion, les stimuli bien reconnus des stimuli moins bien reconnus, et utilisent les premiers pour calculer des profils prototypiques de certaines émotions.

choix contient un élément émotionnellement neutre (*naming*), des émotions (*angry, frightened, content, sad*) et des éléments typiquement liées à l'interaction sociale (*commanding, pleading, astonished, admiring, scornful*). Chaque connotation était définie par une histoire antécédente, que les locuteurs devaient lire avant de prononcer plusieurs fois le mot [saara] sur un ton approprié selon l'histoire.

Une présélection était faite pour écarter des “mauvaises” réalisations. Pour chaque locuteur, trois réalisations de chaque connotation étaient présentées à six juges, à qui on demandait de choisir le meilleur des trois exemples pour la catégorie visée, et d'indiquer si le meilleur exemple était “très mauvais” par rapport à la catégorie. Pour quatre locuteurs, plus de 11 stimuli sur 60 étaient “très mauvais”. Les stimuli de ces locuteurs étaient écartés de l'étude. Pour le test de perception et l'analyse acoustique, les “meilleures” réalisations (une par catégorie) des 12 locuteurs restants, sept femmes et cinq hommes, étaient utilisées.

73 personnes en six groupes participaient au test de perception. Chaque stimulus était présenté quatre fois, suivi d'une pause de 5 à 6 secondes avant le stimulus suivant. Les énoncés des locuteurs masculins n'étaient pas mélangés aux énoncés des locutrices. Les tests duraient 19 min. pour les stimuli des locutrices et 15 min. pour les locuteurs. La tâche des auditeurs était de choisir une parmi dix catégories. L'amplitude des stimuli était normalisée pour le test de perception, pour que les auditeurs ne soient pas dérangés par des stimuli très forts. Les auteurs remarquent que la perception du signal n'est pas altérée par cela : “*The audible cues of the signal waveform enable the listener to judge correctly the speaker's effort to sound pressure irrespective of the playback intensity of a recorded signal*” (p. 1855)

Les résultats sont les suivants. Le taux moyen de reconnaissance est de 50 % (42 % pour les locuteurs, 56 % pour les locutrices). Le taux de reconnaissance varie beaucoup avec la catégorie : entre 70 % et 75 % d'identification “correcte” pour *commanding, angry, et astonished*, mais seulement entre 26 % et 37 % pour *content, admiring, sad, et pleading*. Ces dernières quatre catégories étaient mieux reconnaissable dans les énoncés des locutrices (57 % de bonnes réponses) que dans les énoncés des locuteurs (35 % de bonnes réponses). Aucun locuteur n'était arrivé à communiquer toutes les connotations. Par contre, toutes les catégories contenaient des énoncés dont le jugement était partagé par plus de 70 % des juges.

La présentation graphique que nous avons déjà appliquée à la matrice de confusion de Banse & Scherer (voir Figure 2), appliquée à la matrice de confusion de Leinonen et al. (1997) fait sortir quelques points intéressants (voir Figure 3). Tout d'abord, il est remarquable que *commanding* a une force d'attraction exceptionnelle : la somme des pourcentages dans la colonne “catégorie choisie” de *commanding* est de 149. Seule la catégorie neutre *naming*, qui semble fonctionner par choix “par défaut” en cas de doute, se rapproche de cette valeur. Les catégories *angry, scornful, pleading, sad, et astonished* semblent être relativement stable du point de vue de leur force d'attraction, tandis que *frightened, admiring, et content* ne sont pas des choix attractifs. Quant aux confusions, il existe deux paires remarquables : *angry/commanding* et *pleading/sad* sont mutuellement confondus, indiquant ainsi une proximité perceptive forte. Dans la terminologie symptôme/signal (présentée en 0), une interprétation possible serait que pour le signal *commanding (pleading)*, l'émetteur se base sur le symptôme *angry (sad)*. Les catégories *frightened* et *scornful*, peu confondues avec d'autres catégories, semblent avoir des indicateurs

acoustiques relativement spécifiques. La confusion entre *content* et *astonished* peut être interprétée soit comme une parenté acoustique, soit par une parenté sémantique prêtant à confusion au moment du choix dans le test de perception, soit encore par une production d'étonnement positif pendant l'établissement du corpus. La production ambiguë serait aussi une explication possible pour la confusion unilatérale de *naming* avec *commanding* : le scénario pour *naming* met le locuteur dans la situation d'un président donnant la parole à Sarah. Une dernière observation est celle de l'absence de confusions parmi les "émotions" *angry*, *frightened*, *sad*, et *content* entre elles ou parmi les "motivations" entre elles.

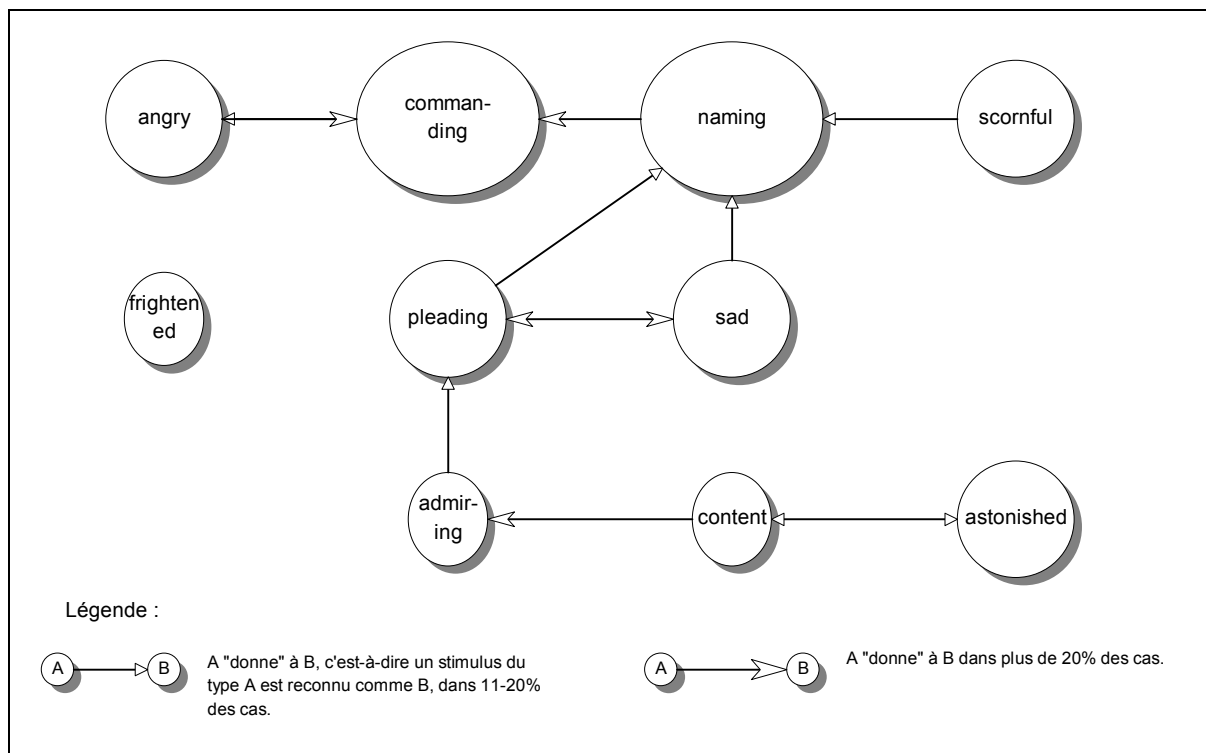


Figure 3. Une présentation graphique de la matrice de confusion de Leinonen et al. (1997).

Les points représentent les 10 catégories du test de perception. La taille d'un point représente le nombre de fois que la catégorie associée a été choisie, indépendamment de la justesse des choix. Ce nombre se calcule dans la matrice de confusion comme la somme des pourcentages dans chaque colonne "catégorie choisie". Cette valeur est un indicateur de l'attrance de la catégorie dans les choix des locuteurs. Un point de taille moyenne indique une valeur entre 80 et 120, un petit point une valeur au-dessous de 80, et un grand point une valeur au-dessus de 120. Les flèches entre points représentent des confusions entre catégories.

Pour l'analyse acoustique, les paramètres suivants étaient analysés : durée de [s], [aa], et [ra] ; moyenne de F0 dans les segments [aa] et [a], registre dans le segment [aa], énergie de [aa], et différences d'énergie entre [s] et [aa] et entre [aa] et [ra]. Des courbes d'énergie des signaux [saara], des contours de F0 pendant le segment [aa] et des modèles spectraux pour [aa] étaient évaluées visuellement. Les locuteurs et les locutrices étaient des groupes relativement homogènes par rapport aux paramètres acoustiques : seulement pour la durée, des différences interindividuelles significatives ont été trouvées.

Dans un *Sammon plot* des paramètres acoustiques énergie, moyenne de F0, registre, et durée du segment [aa], il est possible d'identifier deux zones bien distinctes pour les catégories " *soft and low* " (*naming, sad, pleading, content, admiring, scornful*) et " *high and loud* " (*commanding, angry, frightened, astonished*). Des contours stéréotypes de F0 ont été trouvés pour *astonished, pleading, et scornful*. Les informations spectrales ont été analysées avec la carte auto-organisée (*self-organized map*) de Kohonen. Cette carte présente de manière visuelle le regroupement des connotations en domaines spectraux et les relations entre les groupes. Les modèles *commanding* montrent relativement plus d'énergie entre 1 et 4 kHz que les modèles *naming* ; ce déplacement de l'énergie est encore plus prononcé pour les modèles *angry*. Les modèles *sad* sont caractérisés par une emphase relative des basses par rapport aux hautes fréquences. Le modèle *admiring* montre un spectre plutôt plat et bruité, ce qui est dû à une phonation aspirée (*breathy*). La comparaison des modèles *angry* et *frightened* suggère que les signaux de *frightened* avaient relativement plus d'énergie à la fréquence fondamentale et moins d'énergie dans les formants autour de 1 à 4 kHz.

Cette expérience a montré que 10 connotations émotionnelles/motivationnelles peuvent être communiquées dans un seul mot. Les taux de reconnaissance moyen et différentiels sont très proches de ceux trouvés par Banse & Scherer (1996). L'article de Leinonen et al. est aussi particulièrement intéressant à cause des moyens de présentation des analyses statistiques, comme le *Sammon plot* et la carte spectrale auto-organisée, entre autres.

5. Le rire et le sourire

5.1. Le sourire audible

Tartter (1980) a montré qu'une expression faciale souriante a des effets audibles sur le signal de parole. Cet effet a une explication articulatoire : le sourire raccourcit le conduit vocal et élargit l'ouverture de la bouche, ce qui devrait engendrer une augmentation des fréquences des formants et l'amplitude.

Dans l'expérience de Tartter (1980), six locuteurs américains entraînés, trois hommes et trois femmes, ont produit 25 syllabes sans sens et 4 phrases pleines, dans deux conditions : d'une part, avec une expression faciale neutre, d'autre part, avec une expression faciale souriante, mais sans essayer d'exprimer de la joie. Dans un test de perception, ces énoncés ont été présentés par paires souriant/non souriant, chaque paire dix fois, en ordre randomisé, mais sans mélanger les locuteurs. Dans une première condition, les juges étaient renseignés comment les stimuli avaient été produits et devaient choisir l'item ayant été produit avec un sourire. Dans une deuxième condition, la tâche consistait à choisir l'item plus "gai", mais sans savoir comment les énoncés ont été produits ; et dans une troisième condition, également sans information quant à la production des énoncés, les juges devaient choisir l'item plus "triste". Dans chaque condition, 12 étudiants sans entraînement particulier participaient au test.

Les résultats montrent clairement que le sourire est audible et qu'un énoncé souriant est interprété comme plus "gai" et moins "triste" que l'énoncé neutre correspondant. Les résultats varient selon le locuteur. En condition 1, le sourire a été correctement identifié dans

70.8 % des cas, avec une différence significative selon le locuteur (de 63.0 % à 82.2 %). En condition 2, l'item souriant a été reconnu comme plus gai dans 63.4 % des cas (selon le locuteur de 52.8 % à 76.5 %). En condition 3, le taux de reconnaissance moyen était de 55.5 %. Seulement pour quatre des six locuteurs, l'item neutre a été choisi comme triste (selon le locuteur de 57.7 % à 77.8 %) ; pour un locuteur, les juges ont répondu au hasard, et pour un locuteur, l'item souriant a été choisi comme plus triste de manière fiable (dans 72.6 % des cas). Ces résultats montrent que la tâche de sélection formulée en termes d'expression est plus facile que formulée en termes d'émotion.

Trois locuteurs avaient globalement les meilleurs taux d'identification à travers les conditions. Dans une analyse acoustique, des résultats cohérents ont été trouvés pour ces trois "meilleurs" locuteurs : les énoncés avec sourire ont montré *une augmentation des fréquences des trois premiers formants et de F0 et une augmentation de l'amplitude*, vis-à-vis des énoncés sans sourire. Aucun effet significatif de la durée a été trouvé. Les trois autres locuteurs, avec des taux d'identification inférieurs, différaient des autres en au moins un de ces paramètres.

Tartter & Braun (1994) ont tenté de reproduire les résultats de l'expérience de Tartter (1980). Ils ont fait produire uniquement des syllabes [hVd], avec 10 voyelles différentes, à six locuteurs américains. Dans un test de perception avec six juges étudiants, américains également, la condition 2 de Tartter (1980) a été reproduite, sauf que cette fois-ci, les locuteurs étaient mélangés. Le taux moyen de reconnaissance des syllabes avec sourire comme étant plus gai est de 54.4 %, ce qui est inférieur au taux de Tartter (1980), mais significativement mieux que le hasard. Pour un locuteur, les syllabes avec sourire ont été mieux reconnues que pour les autres locuteurs, de manière significative. Les mesures acoustiques ont trouvé, pour le sourire, une augmentation significative de F2, ainsi qu'une augmentation de la durée des syllabes. Aucune augmentation de F0 n'a été trouvée.

Ces expériences montrent que même pour des monosyllabes sans sens, sans contexte quelconque, et sans entraînement, rien qu'à partir du signal audio, il est possible de distinguer de la parole produite avec un sourire et sans sourire. Ceci semble plus facile si les locuteurs ne sont pas mélangés. Il nous semble important de remarquer que dans les deux expériences, les sourires étaient produits de manière purement mécanique, avec la consigne expresse donnée aux locuteurs de ne pas exprimer de la joie, mais de seulement tendre les muscles faciaux pour produire une expression faciale souriante. Il ne s'agit donc pas ici d'études sur l'expression émotionnelle naturelle, qui peut impliquer des changements de timbre et d'intonation, mais d'études sur la correspondance entre l'expression faciale et des effets perceptifs et acoustiques au niveau du signal de parole.

5.2. Le sourire de Duchenne – marque d'amusement

Ekman et al. (1990) ont trouvé une différence entre un "sourire de Duchenne", lié à l'amusement et à la joie, et d'autres sourires, sans ce rapport. L'idée d'un sourire involontaire de joie, qui se distinguerait d'un sourire volontaire de politesse par la contraction de *l'orbicularis oculi*, remonte à Duchenne de Boulogne, 1862. Ekman et al. (1990) ont saisi le phénomène avec les méthodes de la science contemporaine.

L'expression faciale, l'électroencéphalographie (EEG), et le compte-rendu de l'état émotionnel subjectif ont été enregistrés pour 34 femmes droitières en train de regarder des films plaisants et déplaisants. Pendant que les sujets regardaient individuellement deux films plaisants et deux films déplaisants, leur EEG était enregistré. Après chaque film, qui durait environ 90 s, les sujets devaient évaluer leur propre état émotionnel : une à une, 11 étiquettes émotionnelles étaient proposées, et les sujets devaient indiquer l'intensité de l'émotion correspondante sur une échelle unipolaire de 0 à 8, une valeur plus grande correspondant à un vécu émotionnel plus intense. L'expression faciale était enregistrée par une caméra cachée, et codée en termes de muscles tendus avec le *Facial Action Coding System (FACS)* d'Ekman et Friesen (1976, 1978), par des codeurs expérimentés qui ne connaissaient pas les hypothèses de l'expérience. Ensuite, les séquences où, en plus du grand zygomatique, la partie extérieure de l'orbiculaire de l'œil (*orbicularis oculi, pars lateralis*) était contracté, étaient classées comme "sourire de Duchenne" ; si le grand zygomatique, mais non pas l'orbiculaire de l'œil était contracté, la séquence était classée comme "autre sourire". La durée totale de présence de chaque type de sourire pendant un film était ensuite calculée comme mesure de présence de ce type de sourire.

Les résultats montrent la différence entre le sourire de Duchenne et d'autres sourires à différents niveaux. En ce qui concerne le type de film, les sourires de Duchenne apparaissent significativement plus souvent pendant les films positifs que pendant les films négatifs, tandis qu'il n'y a pas de différence significative pour les autres sourires. Quant à l'expérience subjective, seuls les sourires de Duchenne sont corrélés à des sentiments d'amusement et de joie. En plus, la durée de sourires de Duchenne pendant un film est en corrélation significative avec l'intensité d'amusement et de joie, dans le sens suivant : celui des deux films positifs pendant lequel une durée de sourires de Duchenne plus longue est mesurée suscite aussi une auto-évaluation plus forte en joie et en amusement. De nouveau, cette distinction n'est pas possible à partir des autres sourires. Finalement, au niveau de l'EEG, il existe des différences significatives. Les sourires de Duchenne sont associés à plus d'activation pariétale et temporale antérieure du côté gauche en comparaison avec d'autres sourires.

Ce travail d'Ekman et al. nous apporte ici deux renseignements intéressants. D'une part, l'existence d'une distinction entre différents types de sourires, c'est-à-dire au niveau de l'expression émotionnelle, est liée à des différences d'activité cérébrale. Ceci laisse supposer qu'il pourrait également exister une différence au niveau de l'expression vocale. D'autre part, tous les sourires, de Duchenne ou non, apparaissent dans une situation non sociale, solitaire, devant un écran vidéo. Ceci va à l'encontre des positions qui voient le sourire comme une expression purement sociale. La présente expérience montre bien qu'il existe un type de sourire en corrélation avec des états d'expérience émotionnelle subjective, qui est donc du type SYMPTOME et non pas du type SIGNAL (voir 3.1. La tripartition d'Ohala, 0).

5.3. Le rire

Provine (1996), à partir d'enregistrements de 1200 rires d'hommes et femmes américains en contexte naturel, décrit la structure acoustique et la fonction sociale du rire.

5.3.1. Les caractéristiques acoustiques du rire

Provine (1996) décrit la structure acoustique du rire humain. Un rire est une série de segments ressemblant à des voyelles, chacun d'une durée d'environ 75 ms, avec des intervalles régulières de 210 ms entre deux segments. Il n'existe pas de voyelle spécifique pour le rire, mais lors d'un rire, la voyelle reste la même. Le F0 du rire d'une femme est situé vers 502 Hz, celui d'un homme vers 276 Hz. Si les sons expiratoires entre les segments du rire sont remplacés, dans le signal, par des intervalles de silence, le rire est pourtant reconnu comme normal : l'intervalle de temps entre deux segments porte de l'information, mais le son expiratoire n'en porte pas. Comme la parole, le rire est caractérisé par un *decrecendo* naturel, c'est-à-dire l'amplitude des derniers segments dans un rire est inférieure à celle des premiers, ce qui s'explique probablement par la diminution de la pression subglottique au cours de l'expiration.

Le rire apparaît en règle générale *après* une phrase de parole et ne l'interrompt pas. Il est donc probable qu'un mécanisme neurologique règle l'emplacement du rire dans la parole, la priorité d'accès au canal de vocalisation unique étant donné à la parole. Néanmoins, il existe une forme hybride entre parole et rire, "*laugh-speak*", où le locuteur rit pendant qu'il parle, et que Provine qualifie de "*consciously controlled*" (p. 42).

Il est à noter que le rire humain consiste de segments produits par l'interruption du courant d'air lors d'une seule expiration. Chez les chimpanzés, il existe une vocalisation que Provine appelle le "rire" des chimpanzés. Comme le rire humain, elle est composée d'une suite de sons et de pauses, mais les chimpanzés produisent un seul son par expiration et un seul par inspiration. Ils ne sont pas capable de moduler le courant d'air expiratoire.

5.3.2. Le rôle social du rire

Provine (1996) voit le rôle du rire comme une fonction sociale et non pas comme une expression égocentrique d'émotion. Dans les rires qu'il a observé, moins de 20 % des occurrences du rire étaient une réponse à un effort formel humoristique. La grande partie du rire conversationnel suit des remarques banales, et semble être dû à une ambiance émotionnelle positive, folâtre, et un sentiment de groupe, et non pas à de la comédie. Provine en conclut qu'une recherche qui se focalise seulement sur la réponse des sujets à des plaisanteries ne saisit qu'un petit sous-ensemble du rire.

Il a également observé une différence de la quantité du rire selon le sexe. En règle générale, les femmes rient plus que les hommes. Il s'y ajoute l'observation étonnante qu'en moyenne, le locuteur rit 46 % plus souvent que le destinataire. Ainsi, une locutrice rit 127 % plus que son interlocuteur masculin, et un locuteur rit 7 % moins que son interlocutrice.

Provine formule l'hypothèse que le rire peut être un signal de dominance/soumission ou d'acceptation/rejet. Pour en savoir plus, il faudrait étudier le rire dans des groupes de personnes qui diffèrent en position sociale et en sexe.

Un aspect du rire humain au sujet de son rôle social est le fait qu'il est contagieux. Le simple fait d'entendre un rire incite des sujets à rire ou au moins à sourire. Provine spécule que des humains pourraient posséder des détecteurs auditifs, des circuits neuraux réagissant uniquement à cette

vocalisation typique pour l'espèce humaine. Ensuite, ces détecteurs stimuleraient les circuits neuraux générateurs de l'action stéréotypée qu'est le rire.

5.4. Neurophysiologie du sourire et du rire

5.4.1. Un centre déclencheur du rire : la SMA

Fried et al. (1998) ont détecté que la stimulation d'une partie antérieure de la SMA (*supplementary motor area*, aire motrice supplémentaire) humaine peut engendrer le rire et le sentiment de gaieté correspondant. La surface corticale du lobe frontal gauche d'une fille épileptique de 16 ans a été examinée à l'aide de stimulations électriques, pour localiser le centre déclencheur des attaques dans le but de sa résection chirurgicale. Les attaques de la patiente n'étaient jamais accompagnées d'un rire, ce qui laisse supposer un fonctionnement normal du cerveau en ce qui concerne le rire. Dans une petite aire de 2 cm x 2 cm dans la partie antérieure de la SMA, la stimulation déclenchait le rire de manière reproductible, arrêtant toute activité verbale. Ce qui est extraordinaire est l'observation que le rire était accompagné de l'expérience subjective correspondante de gaieté, et que la patiente trouvait à chaque reprise un stimulus externe quelconque auquel elle attribuait son rire (" *you guys are just so funny... standing around* "). Cette observation suggère, selon les auteurs, un lien étroit entre les composantes moteur, affectif, et cognitif du rire. Il ne faut d'ailleurs pas voir la SMA comme le centre déclencheur du rire, mais plutôt comme une composante d'un " *large neuronal network capable of parallel distributed processing, where the entire network is activated as a whole by the stimulation of any of its constituent units* " (Fried et al., 1998).

5.4.2. Du sourire au rire – un continuum neurophysiologique

Une autre observation de Fried et al. (1998), qui pour nous est d'importance, est celle d'un continuum entre le sourire et le rire. Ils remarquent que la durée et l'intensité du rire augmentaient avec le niveau de courant de stimulation. Pour des petits courants, seulement un sourire était présent, tandis que pour des courants plus élevés, un rire contagieux était induit. Ceci suggère que le sourire et le rire sont des phénomènes reliés sur un seul continuum.

5.5. En résumé

Nous avons vu que le sourire est perceptible auditivement, qu'il est possible de distinguer une syllabe énoncée en souriant d'une syllabe avec une expression faciale neutre. Les études de Tartter isolaient cependant la configuration faciale qu'est le sourire de ses causes naturelles, telle que la joie et l'amusement, neutralisant ainsi d'autres effets possibles que pourraient avoir ces causes du sourire sur le signal de parole. Cette possibilité d'expression, en vue d'une récupération perceptuelle éventuelle, n'a pas encore été examinée. Des variations systématiques de paramètres acoustiques pourraient permettre la distinction auditive de différentes expressions impliquant un sourire, telles que l'amusement, et former ainsi un canal parallèle à l'expression émotionnelle par configuration faciale, que décrivent Ekman et al. (1990).

Fried et al. (1998) ont trouvé des indices pour un continuum entre le sourire et le rire. Une expérience visant à susciter un sourire naturel accompagnant de la parole risque donc de parfois déclencher un rire, qui, lui, ne peut pas se superposer à la parole aussi facilement que le sourire. Provine (1996) fournit des informations sur les propriétés acoustiques du rire et l'organisation temporelle entre la parole et le rire. Sans approfondir, il mentionne l'existence du *laugh-speak*, qu'il classifie comme peu naturel ; la plupart du temps, le rire suivrait la parole au lieu de l'interrompre.

L'observation d'un rire, accompagné du sentiment de gaieté approprié, suite à la stimulation d'une partie de la SMA semble nouvelle ; cependant, elle peut être interprétée dans le cadre dessiné par Damasio (1994). En effet, la SMA, bien qu'appartenant au néo-cortex, est avoisiné au système limbique, siège des émotions primaires selon Damasio. La SMA peut donc très bien appartenir au réseau neural effectuant le traitement des émotions. La proximité spatiale suggère qu'elle est directement liée au système limbique, donc aux émotions primaires – innées –, et non pas au cortex préfrontal où sont situées les émotions secondaires – acquises au courant de l'histoire individuelle.

Chapitre 3

QUELQUES PRINCIPES METHODOLOGIQUES

6. Méthodes d'établissement d'un corpus

Beaucoup de travaux sur la parole émotionnelle se basent sur une méthodologie expérimentale. L'établissement du corpus est donc la première étape essentielle à partir de laquelle sont menés des études perceptives ou acoustico-visuelles. Bien entendu se pose à chaque fois le problème de la représentativité du corpus.

6.1. La problématique : la contradiction d'un corpus contrôlé d'émotions naturelles

Le déclenchement des émotions prend son origine dans le système limbique (Ploog, 1979 ; Damasio, 1994, voir 0), et n'est par conséquent pas volontaire. La simulation des émotions semble être une procédure établie socialement qui met en œuvre une partie seulement du processus spontané (Damasio, 1994, voir 0). Il est donc très complexe, voire impossible, d'établir un corpus en demandant à des sujets de se mettre volontairement dans un état émotionnel choisi.

Ainsi, pour l'établissement d'un corpus de parole émotionnelle, les exigences de naturel et de contrôle se contredisent. L'expérimentateur a le choix entre

- l'extrême " naturel " : l'observation de l'expression d'émotions naturelles, en situation (pseudo-) écologique, ce qui implique que l'on n'a pratiquement aucun contrôle sur les émotions dont on enregistre l'expression (Scherer et al., 1984) ;
- l'extrême " contrôle " : on fait simuler exactement ce que l'on veut par des acteurs, ce qui implique que l'on n'a pas affaire à l'expression d'émotions spontanées, et il faut donc se poser à la fois la question de la qualité de la simulation et de sa similitude perceptive avec le spontané (Banse & Scherer, 1996 ; Ladd et al., 1985 ; Leinonen et al., 1997) ;
- tout ce que l'ingéniosité des chercheurs a trouvé comme compromis entre les deux extrêmes (Ekman et al., 1990 ; Gerrards-Hesse et al., 1994).

6.2. Les différentes méthodes d'établissement des corpus

6.2.1. *Emotions naturelles*

Le seul travail expérimental que nous ayons rencontré et qui ait recours à des stimuli issus d'une situation quasi naturelle est celui de Scherer, Ladd & Silverman (1984). Ils ont enregistré des

employés de l'assistance sociale en train de discuter avec des acteurs amateurs, dans un studio d'enregistrement aménagé en tant que bureau (voir 0).

6.2.2. *Emotions simulées par des acteurs*

Deux grandes études publiées récemment (Banse & Scherer, 1996 ; Leinonen et al., 1997) utilisent cette méthode. Elle consiste à faire lire à des acteurs des phrases ou mots soigneusement choisies, avec la consigne d'exprimer telle ou telle émotion. Ainsi les enregistrements permettent des comparaisons entre émotions *ceteris paribus*.

Un problème est la définition précise de la qualité de l'émotion étudiée : Scherer (1986) déplore l'impossibilité de comparer différentes études sur l'expression émotionnelle à cause du manque de détails sur ce que les auteurs entendent par une étiquette comme " joie ", " peur ", " colère " etc. Pour donc définir plus précisément chaque émotion, les deux expériences citées utilisent des histoires cadres, des scénarios définissant de manière assez précise la situation imaginaire dans laquelle l'énonciation est censée avoir lieu. Dans leur article, Leinonen et al. (1997) donnent pour chaque émotion étudiée un résumé de l'histoire cadre utilisée, ce que ne font pas Banse & Scherer (1996) ; par contre, ils ont choisi deux scénarios par émotion dans un corpus de scénarios issu de plusieurs études interculturelles sur les antécédents d'émotions (Scherer et al., 1986 ; Scherer & Wallbott, 1994).

6.2.3. *Emotions suscitées en laboratoire*

Scherer (1986) s'énonce de manière plutôt pessimiste quant à la possibilité de susciter dans le laboratoire des émotions utilisables :

" One of the most serious handicaps in studying emotional expression, or emotion generally, is that ethical concerns as well as strong cultural affect control norms (...) render it virtually impossible to study emotional processes in natural contexts or to experimentally induce strong emotional states in the laboratory. " (Scherer (1986), p. 144)

Pourtant, il existe une multitude de tentatives d'induire des émotions par les moyens les plus divers.

Dans le domaine de la recherche sur l'expression de l'émotion, citons Ekman et al. (1990). Ils ont fait regarder à leurs sujets des films courts, plaisants et déplaisants, pour mesurer leur EEG et, à l'aide d'une caméra vidéo cachée, l'expression de leur visage.

Dans le domaine psychologique plus large, Gerrards-Hesse et al. (1994) ont revu et comparé presque 250 études depuis 1979 concernant l'induction expérimentale de la joie et de la dépression chez des adultes sains. Les études ne sont pas directement concernées par l'expression de l'émotion. La catégorisation proposée par Gerrards-Hesse et al. est la suivante.

1. Méthodes basées sur la génération mentale libre d'états émotionnels.

- Méthode " hypnose " : les sujets entrent en une transe profonde et on leur demande ensuite de se souvenir et d'imaginer une situation de leur propre choix dans laquelle ils se sentaient heureux ou triste.

- Méthode “ imagination ” : les sujets ont pour tâche d’imaginer et d’expérimenter de nouveau des situations ou des événements déjà vécus.
2. Méthodes basées sur la génération mentale guidée d’états émotionnels.
 - Méthode “ Velten ” : cette méthode, nommée selon son inventeur (Velten, 1968), utilise des déclarations se référant à soi-même décrivant des auto-évaluations et des sensations physiques positives ou négatives. Dans la plupart des études, on demande aux sujets d’utiliser les déclarations comme des autosuggestions, c’est-à-dire d’essayer de sentir l’émotion décrite par la déclaration.
 - Méthode “ film/histoire + ”¹³ : les auteurs présentent un film, une histoire ou une description brève d’une situation à leurs sujets et leur demandent d’imaginer la situation et de “ s’impliquer ” dans les émotions suggérées.
 - Méthode “ musique + ” : les sujets écoutent une pièce de musique classique ou moderne suggérant une émotion et sont demandés d’entrer dans l’émotion exprimée par la musique en utilisant la manière qu’ils trouvent la plus efficace.
 3. Méthodes basées sur la présentation de matériel induisant des émotions.
 - Méthode “ film/histoire ” : cette méthode se sert du phénomène que des films qu’on regarde ou des histoires qu’on lit peuvent susciter des émotions. La différence avec la méthode “ film/histoire + ” est qu’on ne demande pas aux sujets d’imaginer les événements en question ou de se sentir concerné par la situation décrite.
 - Méthode “ musique ” : de même, dans cette méthode, on présente une pièce de musique sans mettre l’accent sur son caractère émotionnel.
 - Méthode “ cadeau ” : les chercheurs utilisant cette méthode supposent que la plupart des personnes sont heureux quand ils reçoivent un cadeau inattendu, comme 10 cents ou une barre de chocolat.
 4. Méthodes basées sur la mise en situation émotionnelle en relation avec des besoins.
 - Méthode “ succès/échec ” : touchant au besoin d’accomplissement, cette méthode consiste à faire passer des tests de capacités cognitives aux sujets et de leur donner des fausses évaluations positives ou négatives sur leur performance.
 - Méthode “ interaction sociale ” : se référant aux besoins d’acceptation sociale des sujets, cette méthode expose les sujets à certaines interactions sociales préparées par l’expérimentateur pour induire des états émotionnels.
 5. Méthodes visant la génération d’états physiologiques relevant d’émotions.
 - Méthode “ drogue ” : les sujets prennent une drogue, p. ex. de l’adrénaline, ou un placebo qui leur est présenté comme une drogue induisant une émotion.

¹³ Le “ + ” représente la consigne expresse de ressentir l’émotion suggérée.

- Méthode “ expression faciale ” : selon l’hypothèse de la rétroaction faciale (*facial feedback hypothesis*) de Leventhal (1980), il est possible d’induire une émotion par l’expression faciale correspondante. Par conséquent, dans la méthode “ expression faciale ”, l’expérimentateur demande aux sujets de contracter et détendre différents muscles pour produire un sourire ou pour froncer les sourcils, ce qui devrait induire un état émotionnel positif ou négatif.

La conclusion de Gerrards-Hesse et al. (1994), après avoir analysé statistiquement l’efficacité des différentes méthodes pour induire joie ou dépression, est la suivante : pour l’induction de joie, la méthode “ film/histoire ” et la méthode “ cadeau ” se sont montrées efficaces dans un certain nombre d’études ; pour l’induction de dépression, les méthodes “ imagination ”, “ film/histoire ”, “ succès/échec ” et “ Velten ” semblent efficaces.

Il est à remarquer que des méthodes semblent être utilisées en psychologie (notamment les méthodes “ imagination ” et “ expression faciale ”), dont les psychologues ont apparemment constaté de manière empirique une certaine efficacité, mais dont le fonctionnement commence tout juste à être décrit par des techniques de neurosciences. Dans le chapitre consacré à Damasio (1994), les observations qu’une imagination mentale ou une expression faciale peuvent déclencher la perception d’une émotion étaient des éléments importants et nouveaux, permettant de faire avancer la théorie des structures neurales sous-tendant les émotions (voir 0 et 0).

6.2.4. Stimuli créés artificiellement par resynthèse

Une seule étude a créé des stimuli émotionnels artificiellement, par resynthèse du contour intonatif : Ladd et al. (1985). Se basant sur une théorie phonologique de la prosodie, ils ont fixé des points d’ancrage sur le contour prosodique qui, selon eux, représentaient l’essentiel de l’information prosodique. Pour resynthétiser le contour intonatif, ils ont changé la valeur de F0 à ces points d’ancrage, interpolé automatiquement le contour entre les points et rétabli la microprosodie manuellement. De la sorte, ils ont créé un total de 24 stimuli à partir de 6 énoncés naturels (voir 0).

Une autre application imaginable de la resynthèse de stimuli émotionnels serait de créer un continuum entre deux énoncés émotionnels en contrôlant précisément les variables changeantes. Cela suppose déjà une bonne connaissance des propriétés acoustiques des émotions utilisées. L’intérêt d’une telle expérience serait la question théorique de l’existence d’une frontière perceptive catégorielle entre les deux émotions ou d’une variation graduelle de la perception. Dans le premier cas, on obtiendrait un changement de taux de jugement brusque à la frontière perceptive, dans le deuxième cas, les deux émotions seraient confondues dans une zone relativement large.

6.3. Présélection des énoncés

Pour s’assurer de la qualité des énoncés qui seront utilisés dans l’étude, Banse & Scherer (1996) ont fait juger les énoncés enregistrés par des experts et n’ont retenu que ceux qui ont obtenu un bon jugement. De la sorte, ils ont présenté à des acteurs en formation les 1344 enregistrements audiovisuels qu’ils avaient obtenu par des acteurs professionnels (voir 0). Les enregistrements

ont été présentés dans les conditions audio seul, vidéo seule, et audiovisuel, triés par émotions exprimées. Les experts les jugeaient selon les deux critères authenticité et reconnaissabilité, avec des notes scolaires. Seulement les énoncés ayant obtenu une bonne note ont été retenus, ce qui a réduit le nombre d'énoncés à 280. Ensuite, pour des raisons formelles d'homogénéité de l'expérience, les expérimentateurs ont réduit le nombre d'énoncés à 224.

De même, Leinonen et al. (1997) ont présélectionné leurs stimuli en faisant choisir à six juges parmi trois productions pour chaque connotation et chaque locuteur, avec la possibilité d'indiquer si le meilleur exemple parmi les trois était toujours "très mauvais" comme expression de la connotation visée. Si un locuteur avait trop de jugements "très mauvais", l'ensemble de ses expressions était écarté de l'étude.

7. Test de perception

7.1. Connaissances et meta-connaissances

Les tests de perception ont souvent pour but de mettre en évidence des connaissances manipulées par les sujets durant des processus de perception. Il est classique de demander explicitement aux sujets de manipuler en direct ces connaissances, par exemple en donnant aux sujets une liste d'étiquettes langagières conceptuellement associées aux émotions que l'on veut que le sujet identifie. On le met ainsi dans une tâche de meta-connaissance (connaissance de la connaissance !) : la capacité de *nommer* des émotions met en œuvre un processus linguistique de représentation, externe à la tâche. Si le sujet ne peut pas répondre à la tâche, cela ne signifie pas qu'il ne manipule pas implicitement ces connaissances à l'intérieur du processus de perception. Il faut donc être prudent dans l'interprétation de ce type de test. Ce qui peut néanmoins justifier une telle approche est l'usage social de la capacité de nommer une émotion, comme dans "Tu as l'air triste/content/etc."

Pour éviter de faire appel à des meta-connaissances dans un test de perception, il faudrait donc éviter toute étiquette langagière. Ceci peut être réalisé par un test de préférence ("lequel des deux stimuli préférez-vous ?") ou par un test de comparaison ("est-ce que les deux stimuli sont pareils ou différents ?").

Tous les tests de perception en recherche sur la parole émotionnelle que nous avons rencontrés font appel à des meta-connaissances en utilisant une tâche d'identification langagière. Les bons scores qu'obtiennent ces études prouvent donc l'existence à la fois des connaissances et de la capacité de les expliciter de manière langagière.

7.2. Quelques éléments de tests de perception en recherche sur la parole émotionnelle

7.2.1. Le questionnaire

a) Type de questionnaire

- Identification : dans un questionnaire d'identification, chaque stimulus doit être identifié dans une liste de réponses possibles. Ce sont les études sur les caractéristiques acoustiques que nous avons vu (Banse & Scherer, 1996 ; Leinonen et al., 1997) qui utilisent un tel questionnaire.
- Distinction : dans un questionnaire de distinction, les stimuli sont présentés en paires, et les juges sont demandés lequel des deux est plus X (p. ex. plus "joyeux", plus "souriant"). C'est la méthode utilisée par Tartter (1980) et Tartter & Braun (1994) pour tester si le sourire est audible. La présentation de paires de stimuli se prête donc particulièrement à l'étude de la question d'une distinction perceptive entre deux groupes de stimuli. Il nous semble que, surtout pour un locuteur inconnu, une comparaison est une tâche plus facile qu'une identification, parce que l'auditeur a la possibilité de comparer.
- Echelles de graduation : pour indiquer le degré d'un jugement émotionnel supposé continu, Ekman et al. (1990, voir 0) et Ladd et al. (1985, voir 0) utilisent des échelles de graduation (de 0 à 8). Ekman et al. (1990) mesurent le vécu émotionnel de leurs sujets suite à la présentation de films plaisants et déplaisants par une échelle de graduation unipolaire de 0 à 8 pour chacune de 10 étiquettes émotionnelles. Ladd et al. (1985) utilisent, pour les jugements des états liés à l'excitation physiologique, cinq échelles bipolaires à huit points (détendu/excité, ouvert/trompeur, irrité/content, incertain/arrogant, et indifférent/engagé) ; pour les jugements d'attitudes cognitives, dans une autre session avec les mêmes juges, ils ont utilisé cinq échelles unipolaires à huit points (le degré d'emphase, coopération, contradiction, surprise, et reproche).

b) Choix fermé vs. choix ouvert

Dans la plupart des études que nous avons rencontrées (Banse & Scherer, 1996 ; Ladd et al., 1985 ; Leinonen et al., 1997 ; Tartter, 1980 ; et Tartter & Braun, 1994), les auditeurs sont forcés à choisir parmi les réponses prévues, qui ne contiennent pas de réponse "je ne sais pas", parce que d'une part, cela permet des analyses plus faciles, mais d'autre part, comme le disent Tartter & Braun (1994), "*they were urged to guess since they probably knew more than they thought they did*" (p. 2103). La question soulevée de la conscience qu'ont les sujets de leur savoir est importante ; elle est discutée de manière plus approfondie dans le paragraphe 0.

Une seule des études a proposé un choix ouvert : Scherer et al. (1984, voir 0) ont proposé un questionnaire avec 9 adjectifs émotionnels, dont il était possible de marquer un ou plusieurs, avec un X ou deux si on trouvait que l'adjectif décrivait "extrêmement bien" le stimulus ; en plus, les sujets avaient la possibilité de décrire le stimulus avec leurs propres mots. Cependant, les résultats ont montré que les sujets n'ont que très peu fait usage des libertés offertes : "*As it turned out, subjects usually selected only one or two adjectives per utterance, used two Xs in*

only about 9% of these selections, and generally did not resort to free descriptions ” (Scherer et al., 1984, p. 1348)

c) Neutraliser les effets d'ordre

Une manière simple de neutraliser jusqu'à un certain degré des effets d'ordre éventuels est de proposer les stimuli en ordre inverse à la moitié des juges (Scherer et al., 1984 ; Tartter, 1980) ou, s'il y a deux ensembles de stimuli consécutifs (p. ex. stimuli chuchotés / stimuli parlés normalement), de présenter les ensembles en ordre inverse pour la moitié des juges (Tartter & Braun, 1994, Ladd et al., 1985).

d) Vérifier la fiabilité des juges

Pour tester l'accord entre deux groupes de juges, il est possible de présenter quelques stimuli aux deux groupes (Scherer et al., 1984). De même, pour tester la fiabilité du groupe de juges, Leinonen et al. (1997) ont répété dans le test de perception dix stimuli choisis au hasard, pour comparer les choix faits pour les deux occurrences de chacun de ces stimuli.

7.2.2. Quelques éléments remarquables sur la présentation des stimuli

Ladd et al. (1985) remarquent que les juges s'aperçoivent moins du caractère artificiel des stimuli créés par resynthèse si ceux-ci sont présentés par haut-parleur et non pas par écouteur. Leinonen et al. (1997) ont normalisé l'amplitude des stimuli avant de les présenter, et expliquent que les juges pouvaient quand même reconnaître l'intensité originale des stimuli par les propriétés spectrales du signal (p. 1855).

Tartter (1980) et Tartter & Braun (1994) ont présenté chaque paire de stimuli dix fois, augmentant ainsi la durée du test, mais diminuant le nombre d'auditeurs requis pour atteindre un niveau de signification acceptable¹⁴.

7.2.3. L'exploitation des données

a) Analyse statistique des données

Les méthodes pour calculer la signification sont multiples. Ainsi, Tartter (1980) utilise un “ *two-tailed t test comparing actual performance to chance* ” (p. 25) et “ *a simple (...) analysis of variance* ” (*ibid.*) Tartter & Braun (1994) se servent pour le même but de “ *z approximations to the binomial distribution* ” (p. 2103) Ladd et al. (1985) utilisent, eux aussi, une analyse de la

¹⁴ Pour le nombre de juges, la raison d'en prendre beaucoup (jusqu'à 73, chez Leinonen et al. 1997) est de pouvoir évaluer statistiquement les résultats. Plus le nombre de jugements est grand, plus les résultats sont significatifs. Il est donc possible, pour un grand nombre de juges, de trouver juste un petit écart entre deux valeurs, qui est cependant significatif, c'est-à-dire non explicable par des fluctuations aléatoires. L'idée derrière les tests de signification, issue du calcul des probabilités, est simple : on calcule la probabilité que la distribution des valeurs observée est due au hasard. Si cette probabilité est petite, selon le cas inférieure à 5%, 1% ou 0.1% (niveau de signification $p < 0.05$, $p < 0.01$ ou $p < 0.001$), on peut dire que l'effet est statistiquement significatif, avec le risque d'erreur associé de 5%, 1%, ou 0.1%.

variance, accompagné de “ *d values (... that) are an expression of effect size in standard-deviation units* ” (p. 438) Scherer et al. (1984) mesurent le degré de parallélisme entre deux variables par des “ *correlations* ” (p. 1348). La plus grande variété de tests de signification différents se trouve chez Leinonen et al. (1997). Pour analyser les tests de perception, ils utilisent un “ *chi-square test* ” pour vérifier la fiabilité des juges, un “ *Mann-Whitney U test* ” pour chercher des différences entre juges masculins et féminins, et “ *Spearman’s correlation coefficients* ” pour mesurer les “ *cooccurrences of listener choices in pairs of categories* ”, en précisant que “ *nonparametric tests were used in the present study because populations were not normally distributed and had different variances* ” (p. 1855).

Dans les tests d’identification (Banse & Scherer, 1996, Aubergé et al., 1997 ; Leinonen et al., 1997), les confusions entre classes sont présentées dans une matrice de confusion. L’analyse des confusions donne des indications sur la similitude perceptive entre classes. La présentation graphique que nous avons appliqué aux matrices de confusions de Banse & Scherer (1996, Figure 2, p. 23) et Leinonen et al. (1997, Figure 3, p. 26) permet de visualiser deux types d’informations intéressantes contenues dans une matrice de confusion :

- (1) les confusions importantes entre classes sont visualisées par des flèches entre classes ; deux classes reliées par une flèche sont donc perceptivement proches ;
- (2) le pouvoir d’attraction perceptive d’une classe est visualisé par la taille du point représentant la classe. Des préférences perceptives sont ainsi facilement repérables.

b) *Trouver des réalisations type*

Une possibilité intéressante est celle d’utiliser les exemples bien reconnus dans un test de perception comme réalisations type, permettant de calculer des prototypes acoustiques. Ainsi, Johnstone et al. (1995) réinterprètent les résultats obtenus par Banse & Scherer (1996), en séparant les exemples bien reconnus de ceux mal reconnus pour chaque émotion. A partir des exemples bien reconnus, ils établissent ensuite des prototypes acoustiques pour les émotions correspondantes. De même, Leinonen et al. (1997) analysent les propriétés spectrales pour les différentes émotions à partir d’énoncés reconnus par au moins 50% des juges.

ANALYSE DU CORPUS “ SOURIRE ”

8. L’expression prosodique de l’amusement

Comme premier pas vers une synthèse audiovisuelle d’amusement dans la parole, nous avons établi et analysé un corpus audiovisuel de parole amusée. Le but principal de l’étude est la mise en évidence d’une expression prosodique de l’amusement spontané. Une question centrale est l’apport respectif des canaux acoustique et visuel à la perception d’amusement dans une phrase de parole ; plus particulièrement, dans le canal acoustique, l’expression de l’amusement se résume-t-elle aux effets acoustiques du sourire (Tartter, 1980, Tartter & Braun, 1994), ou est-ce que l’amusement s’exprime par une prosodie particulière ? Une autre question adressée est celle de la distinction perceptive, d’une part, entre l’amusement spontané et l’expression volontaire de l’amusement (joué ou simulé), et d’autre part, entre un sourire d’amusement et un sourire social (Damasio, 1994 ; Ohala, 1996).

8.1. Fondements théoriques de l’expérience

8.1.1. Amusement, énonciation souriante, et énonciation neutre

Notre hypothèse principale est que l’expression de l’amusement se fait selon deux canaux :

- visuel : sur les lèvres, l’amusement s’exprime par un sourire, accompagné éventuellement de changements dans le reste du visage, surtout autour des yeux (Ekman et al., 1990) ;
- acoustique : à part les effets audibles du sourire (Tartter, 1980 ; Tartter & Braun, 1994), l’amusement peut s’exprimer par des changements acoustiques, notamment dans la prosodie.

Les résultats de Tartter (1980) et de Tartter & Braun (1994) montrent que le sourire (intentionné, non accompagné chez le locuteur d’une émotion particulière) s’entend dans la voix dans des logatomes pour des locuteurs américains, et qu’il est interprété comme une marque de gaieté. Le sourire, raccourcissant le conduit vocal et élargissant son ouverture, a des effets mesurables sur le signal acoustique (notamment une augmentation de l’amplitude et des fréquences des formants, voir 0).

Par contre, à notre connaissance, l'effet audible du sourire "mécanique" n'a encore jamais été opposé à une expression amusée spontanée, qui pourrait contenir notamment des changements prosodiques en sus des effets du sourire.

On obtient deux paires minimales :

⇒ les mêmes énoncés lus avec amusement spontané vs. lus de manière émotionnellement neutre ;

⇒ les mêmes énoncés lus avec amusement spontané vs. avec un sourire mécanique.

La récupération perceptive de la première paire (amusement spontané / neutre) sera testée en conditions audio seul, vidéo seule, et audiovisuel.

L'étude de la deuxième paire (amusement spontané / sourire mécanique) porte sur les effets audibles différents, elle doit donc être présentée en condition audio seul.

8.1.2. *Amusement spontané vs. simulé*

Liberman & Streeter (1978) ont montré la capacité de réitérer les caractéristiques prosodiques d'un énoncé par une boucle de bas niveau sans reconstruction cognitive. Nous voulons tester si cette capacité peut s'appliquer sur les caractéristiques émotionnelles d'un énoncé prosodique, en audiovisuel, soit dans une tâche simple de réitération (séquentielle), soit en réitération synchronisée sur le signal original.

⇒ production spontanée vs. réitérée des énoncés "amusés".

L'effet perceptif de la paire spontané/réitéré est à vérifier dans la condition la plus générale possible, c'est-à-dire en audiovisuel.

8.1.3. *Amusement spontané vs. joué*

Damasio (1994) décrit une boucle de simulation des émotions (voir 0), qui permettrait de percevoir des émotions sans les causes corporelles. Une différence d'état intérieur du locuteur entre émotion réelle et émotion simulée peut entraîner des effets différents au niveau de l'expression.

Ekman et al. (1990) avancent l'idée d'une expression différente dans le visage entre le "sourire de Duchenne", correspondant à l'amusement, et d'autres sourires, non liés à l'amusement, les deux types de sourires occurring chez lui spontanément (voir 0). Damasio (1994) pose que le sourire de Duchenne se fait de manière involontaire uniquement ; si on sourit de manière volontaire, *l'orbicularis oculi* ne se contracte pas (voir 0).

À notre connaissance, une différence éventuelle au niveau acoustique entre une expression émotionnelle spontanée, involontaire, et une expression émotionnelle volontaire, intentionnée, n'a pas encore été cherchée. Pour ce faire, nous construisons une paire minimale :

⇒ production spontanée vs. jouée (acteur).

De même, l'effet perceptif de la paire spontané/joué est à vérifier dans la condition la plus générale possible, c'est-à-dire en audiovisuel.

8.1.4. *Sourire amusé vs. sourire social*

Ohala (1996) distingue entre des expressions émotionnelles du type “symptôme” (expressions non intentionnées d'un état intérieur), et des expressions émotionnelles du type “signal” (expressions intentionnées pour agir sur l'autre). Ohala base sa distinction sur des observations éthologiques (voir 0). Cependant, il ne formule aucune hypothèse sur une différence entre les productions issus des deux types d'expression.

Pour chercher une différence entre les productions, nous construisons une paire minimale :

⇒ production acteur d'amusement vs. acteur de séduction.

8.1.5. *Hypothèses*

- (1) L'expression de l'amusement se fait par les canaux de l'expression faciale et de l'expression vocale. Les effets visuels et acoustiques sont récupérables perceptivement ensemble et séparément.
- (2) Les effets acoustiques de l'expression vocale de l'amusement comportent des effets prosodiques en sus des effets acoustiques du sourire tels que les décrivent Tartter (1980) et Tartter & Braun (1994). Ces effets prosodiques de l'amusement permettent de distinguer perceptivement en audio seul un énoncé produit avec de l'amusement spontané d'un énoncé produit avec un sourire mécanique à la Tartter.
- (3) L'expression spontanée et l'expression réitérée de l'amusement sont perceptivement distinguées en condition audiovisuelle.
- (4) L'expression spontanée et l'expression jouée (acteur) de l'amusement sont perceptivement distinguées en condition audiovisuelle, notamment par le plissement de la peau au niveau des yeux dû à la contraction de l'orbiculaire de l'œil. Les expressions d'acteur peuvent ainsi être reconnues comme non spontanées.
- (5) L'expression jouée (acteur) de l'amusement est perceptivement distincte de l'expression jouée (acteur) de la séduction et d'une énonciation avec un sourire mécanique, en condition audiovisuelle.

8.2. **Etablissement du corpus**

L'enjeu principal dans l'établissement du corpus était l'obtention des phrases de parole contenant l'expression spontanée d'amusement. Pour cela, nous avons construit la tâche de mise en situation suivante. Chaque locuteur (4 en tout) était en chambre sourde, avec la consigne de lire une phrase proposée sur un écran d'ordinateur, puis de la répéter en “ma-

ma-ma ”. Chaque phrase était surmonté d’un portrait de personne, afin d’habituer le sujet à la présentation d’une image associée, le prétexte choisi étant que la phrase s’adresse à la personne du portrait. Après une dizaine de phrases, un élément distracteur supposé provoquer la réaction spontanée “amusée” du locuteur était introduit : la projection à l’écran d’un dessin humoristique à la place du portrait, en sus de la phrase à lire et à répéter en “ma-ma-ma”, la phrase étant le commentaire ironique du dessin. Les dessins étaient d’abord séparés par des phrases de consigne, puis de plus en plus rapprochées. Pour obtenir un effet crescendo, les premiers dessins étaient de l’humour ironique anglais (des dessins de Gary Larson, avec commentaires traduits en français), suivis de dessins issus de “Le Chat” de Geluck, pour finir sur un dessin plus “appuyé”, dans lequel nous avons inséré par montage la tête d’un des locuteurs.

Nous avons décidé de filmer le locuteur de face et de profil, parce que l’image profil renseigne sur l’étirement et le relèvement des coins des lèvres, mouvement qu’on perdrait avec l’image de face seule, et qui est pourtant très prononcé pour le sourire.

Pour l’analyse automatique de la configuration faciale, il aurait été bien de pouvoir utiliser le système d’analyse automatique de la position des lèvres qui existe à l’ICP. Ce système a de hautes exigences quant à la qualité du signal vidéo : les lèvres doivent être peintes en bleu, *éclairées intensément*, et filmées avec une caméra de haute précision. La couleur bleue est utilisée parce qu’elle est autrement absente dans le visage (à part éventuellement la couleur des yeux), ce qui permet de remplacer juste le bleu sur les lèvres par un noir parfait, et de calculer par la suite automatiquement les paramètres caractérisant la configuration des lèvres. Cependant, comme nous voulions aussi filmer les yeux, l’utilisation d’un éclairage trop fort s’interdisait, ce qui rendait peu probable de pouvoir utiliser le système d’analyse automatique.

8.2.1. La préparation des sessions d’enregistrement

Les dessins humoristiques de Gary Larson ainsi que le dessin “appuyé” ont été trouvés sur Internet ; ceux de Geluck (1997) ont été numérisés avec un scanner noir et blanc. Tous les dessins ont été retouchés et coloriés, pour augmenter leur signification humoristique. Les portraits ont été trouvés sur le serveur WWW de l’ICP. Leur taille a été adaptée pour correspondre à peu près à celle des dessins.

Pour la présentation sur écran des images et des phrases, un programme Supercard a été écrit, affichant une phrase et l’image associée à chaque fois que l’utilisateur cliquait avec la souris sur un bouton.

Le matériel d’enregistrement était le suivant.

Enregistrement audio :

- microphone dynamique Sony F-PV 250 ;
- préamplificateur ICP ;
- le son était enregistré sur la piste son de la bande vidéo.

Enregistrement vidéo :

- deux caméras JVC KY-15E pour face et profil ;
- unité de contrôle pour chaque caméra JVC RM-P200 ;
- chroma keyer Sony CRK-2000P pour remplacer le maquillage bleu par un noir parfait¹⁵ ;
- table de mixage vidéo “ digital WJ-AVES ” pour mixer les enregistrements face et profil sur une seule piste vidéo ;
- time coder Sony FCG-700 ;
- magnétoscope Betacam SP Sony UVW 1400P ;
- deux cassettes vidéo MLSP 60 min ;
- moniteur de contrôle Sony PVM-14400M.

Eclairage :

- éclairage indirect Mandarine 500 W avec parapluie pour diffraction.

Cette configuration nous a permis d’enregistrer en audio et en vidéo, de face et de profil, avec 50 demi-images par seconde ; de remplacer la couleur bleue du maquillage par un noir complet ; de combiner les deux images en une seule, profil à gauche, face à droite ; et d’ajouter un *time code* avec un numéro unique pour chaque image.

Pour susciter l’amusement, des dessins devaient être affichés sur écran ordinateur. Pour les enregistrements des énoncés réitérés, il était nécessaire de montrer et de faire écouter les enregistrements amusés spontanés établis antérieurement. Ceci a pu être réalisé en raccordant le signal vidéo du lecteur Betacam à l’entrée vidéo du PowerMac. Pour écouter le signal audio, le locuteur disposait d’un écouteur.

Affichage :

- PowerMac 7500/100 avec carte d’acquisition vidéo ;
- écran 17" ProNitron 85.17 ;
- lecteur vidéo Betacam SP Sony UVW 1600P.

Ecouteur :

- écouteur une oreille.

¹⁵ Comme nous l’avons appris plus tard, les spécialistes de la parole visuelle à l’ICP utilisent le chroma keyer uniquement pour vérifier la qualité du maquillage ; pour remplacer le bleu par un noir, ils disposent d’un programme sous ordinateur Silicon Graphics plus performant que le chroma keyer analogique.

8.2.2. *Les locuteurs*

Quatre locuteurs masculins ont été enregistrés : trois locuteurs professionnels de l'ICP (YM, JS, PB) dont les caractéristiques articulatoires audiovisuels sont bien connus, et un locuteur naïf.

Le déroulement des sessions d'enregistrement nous a montré que le choix de prendre des locuteurs professionnels présente un inconvénient lié à la tâche. Chaque locuteur était concentré et contrôlait par souci professionnel au maximum son envie de sourire afin de respecter la consigne lourde imposée. Un locuteur remarquait après la session d'enregistrement "mise en situation" qu'il s'était contrôlé pendant la lecture et que maintenant, les phrases le feraient rire.

A la suite de ces résultats maigres obtenus avec les locuteurs professionnels, nous avons décidé d'enregistrer un locuteur naïf (DV), susceptible de moins se contrôler. Celui-ci s'est avéré très facile à faire sourire pendant la mise en situation ; c'est donc lui qui a contribué le plus à notre corpus.

8.2.3. *Le protocole d'enregistrement*

En vue des mesures manuelles envisagées et d'une éventuelle possibilité d'analyse automatique, les sujets ont été maquillés. Les lèvres ont été maquillées en bleu ; des croix bleues ont été peintes sur les joues, en suivant à peu près le grand zygomatique, et à côté des yeux, pour les plissements de la peau attendus pour les sourires de Duchenne ; une croix fixe de référence a été dessinée sur le nez et sur le menton¹⁶.

Le locuteur YM, passant le premier, n'a pas produit tous les expressions que nous avons prévus (ni "neutre" ni "sourire mécanique") Comme il a systématiquement produit chaque phrase deux fois, nous avons pris, pour les phrases avec amusement, la première de chaque paire comme énonciation amusée et la deuxième comme énonciation neutre. A la suite de cet oubli, nous avons préparé un protocole "aide-mémoire" pour fixer l'ordre dans lequel nous ferions les manipulations, pour n'en oublier aucune, et aussi pour nous rappeler les changements de câblage technique nécessaires entre les manipulations. Ce protocole (voir Annexe) a été respecté pour les locuteurs JS, PB, et DV, sauf pour PB, où l'énonciation avec un sourire social venait à la fin, après la réitération.

La qualité de l'enregistrement audio est critiquable sur deux points : d'une part, pour certaines parties, le signal est légèrement saturé sur la bande ; d'autre part, nous avons enregistré aussi le bruit du ventilateur de l'ordinateur qui se trouvait dans la chambre sourde.

a) *L'enregistrement de l'amusement*

La première manipulation était la mise en situation pour susciter l'amusement spontané chez un locuteur. Comme la consigne était de lire les phrases proposées sur écran et de les répéter

¹⁶ Nous n'avons pas pensé à étalonner l'image avec une règle posée au départ à côté de la tête du locuteur.

en “ ma-ma-ma ”, aussi les phrases de distraction, supposées provoquer de l’amusement, étaient répétées en “ ma-ma-ma ”. L’intérêt de cela était d’obtenir des expressions prosodiques de l’amusement sans interférences lexicales. Ce but n’a été atteint que pour le locuteur DV : il était le seul à sembler amusé pendant la répétition en “ ma-ma-ma ”.

Après la session de mise en situation, nous avons regardé l’enregistrement avec le locuteur et désigné avec lui les énoncés dans lesquels il avait exprimé de l’amusement.

b) *Productions “ acteur ”, “ sourire social ”, “ sourire mécanique ”, et “ neutre ”*

Les phrases sélectionnées ont été écrites sur une feuille de papier, et lues par le locuteur avec l’intention de produire l’expression demandée. Pour chaque expression, toutes les phrases ont été lues avant de passer à une autre expression.

Pour la production “ acteur amusé ”, nous avons demandé au locuteur de dire les phrases de manière à ce qu’un spectateur croie qu’il était amusé, en se servant des moyens qui lui semblaient bons ; en particulier, nous avons précisé qu’il importait peu s’il ressentait ou non de l’amusement pendant cette production.

Comme contexte des expressions d’un sourire social de séduction, nous avons proposé au locuteur de s’imaginer être en face d’une belle femme qu’il avait envie de séduire, et de prononcer les phrases du corpus comme s’il les disait à cette femme.

Ensuite, il devait produire les mêmes phrases avec un “ sourire mécanique ”, que nous lui avons décrit comme une contraction des muscles faciaux, comme dans une expérience de “ *lip tube* ”, ou comme s’il fallait contracter ces muscles pour se soulager d’une douleur.

Enfin, nous lui avons demandé une énonciation émotionnellement neutre, “ normale ”, des mêmes phrases.

c) *Simulation : doublage de l’enregistrement amusé spontané*

Dans un dernier temps, le locuteur devait tenter de simuler la même expression que celle faite pendant son enregistrement initial amusé. L’image vidéo lui était montrée sur un écran d’ordinateur, et il entendait le son original à l’aide d’un écouteur. Après avoir revu une phrase deux ou trois fois, la bande était arrêtée, et le locuteur devait reproduire la même expression en séquentiel. Nous avons procédé de cette manière pour toutes les phrases choisies comme exprimant de l’amusement.

Après la réitération en séquentiel, il revoyait l’enregistrement d’une phrase deux fois et devait simuler son expression en synchronie avec la troisième répétition.

La tâche de réitération, surtout en synchronie, s’est avérée être une tâche difficile, autant du point de vue technique que du point de vue du locuteur. En ce qui concerne les difficultés techniques, les phrases devaient soit être cherchées dans la bande, ce qui produisait un bruit désagréable dans l’écouteur du locuteur, soit être copiées sur une autre bande, trois fois de

suite ; comme la copie était faite sans banc de montage, les trois répétitions ne commençaient jamais au même endroit, ce qui irritait les locuteurs.

La difficulté essentielle du côté du locuteur consistait, pour la réitération en synchronie, à aligner ses gestes dans le temps sur l'enregistrement. Les conséquences de cette difficulté sont des hésitations peu naturelles, et peu de sourires.

8.2.4. *La numérisation des enregistrements*

La numérisation des enregistrements audiovisuels, en vue d'un test de perception et d'analyses acoustiques et visuelles, était difficile et coûteuse en temps. Le problème était de trouver un ordinateur avec une carte d'acquisition vidéo permettant, d'une part, la numérisation du son à 16 bit, qui était importante pour l'analyse acoustique du son ; et permettant, d'autre part, l'enregistrement des séquences vidéo sous un format QuickTime lisible sous Macintosh, l'équipement utilisé pour le test de perception étant des ordinateurs Macintosh.

Le premier ordinateur que nous avons essayé pour la numérisation ne permettait la numérisation du son qu'à 8 bit. L'alternative, un PC équipé d'une carte d'acquisition vidéo et permettant la numérisation du son à 16 bit, n'était accessible que le soir et le week-end. Par contre, les séquences QuickTime créées avec ce PC n'étaient pas jouées à la bonne vitesse sous Macintosh, et toute tentative de solution de ce problème restait infructueuse.

La dernière alternative, et celle que nous avons fini par prendre, était un Macintosh avec une carte d'acquisition vidéo interne de qualité moyenne. Tout en permettant un son numérisé en 16 bit, ses capacités de numériser l'image étaient moins fortes que pour les deux autres ordinateurs. Ainsi, nous étions forcé à réduire la taille de l'image des 384x288 pixels (demi-écran PAL) souhaitées à 320x240, en préservant 25 images/seconde ; mais même ayant réduit la quantité de données à transférer par seconde de cette manière, la carte était trop faible, perdant ainsi environ une image sur dix.

Les énoncés ont été numérisés, découpés en laissant une demi-seconde avant le début de la parole, et compressés avec la méthode la plus performante (cinepak). Le son a été numérisé en même temps que les images vidéo, à 44100 Hz et sans compression. Toute tentative de réduire le nombre de valeurs par seconde pour le son dégradait sa qualité, en ajoutant un bruit. Ainsi, toutes les séquences audiovisuelles ont été enregistrées avec un son à 44100 Hz, et les fichiers son AIFF ont été exportés également à 44100 Hz / 16 bit.

8.2.5. *Le corpus résultant*

Le résultat est un ensemble de 184 fichiers vidéo en format QuickTime, et de 180 fichiers son en format AIFF. En détail, le corpus contient, pour les différents locuteurs, les phrases suivantes énoncées des manières indiquées.

Explication des tableaux : pour chaque locuteur, les têtes des colonnes indiquent les sept types d'expressions enregistrés, tels qu'ils ont été décrits plus haut : amusement spontané,

amusement joué en tant qu'acteur, sourire social de séduction, sourire mécanique comme expression faciale sans ses causes naturelles, énonciation émotionnellement neutre, répétition des phrases spontanées en séquentiel et en synchrone. Les têtes des lignes indiquent la phrase énoncée ainsi que l'abréviation utilisée pour se référer à la phrase dans l'analyse des résultats du test de perception (voir 0). Un X dans une case indique que le locuteur a prononcé la phrase avec l'expression indiquée.

YM		spontané	acteur	social	mécanique	neutre	réit. séq.	réit. synch.
gynéco	y1	X	X			(X)	X	X
saute-mouton	y2	X	X	X		(X)	X	X
waouh	y3	X	X	X		(X)	X	X
voir vendredi	y4	X	X					X
aJaJaz			X					
ibibiz			X					
ububuz			X					
total 23		4	7	2	0	(3)	3	4

Tableau 1. Les phrases du corpus pour le locuteur entraîné YM.

Les phrases " neutres " sont mises entre parenthèses pour indiquer qu'elles n'ont pas été énoncées dans le même contexte que pour les autres locuteurs (voir 0).

JS		spontané	acteur	social	mécanique	neutre	réit. séq.	réit. synch.
bonne aussi	j1	X	X	X	X	X	X	X
bonne	j2	X	X	X	X	X	X	X
exagerez	j3	X	X	X	X	X	X	X
foot	j4	X	X	X	X	X	X	X
gynéco	j5	X	X	X	X	X	X	X
lumière	j6	X	X	X	X	X	X	X
aJaJaz			X	X		X		
iziziz			X	X		X		
ububuz			X	X		X		
total 51		6	9	9	6	9	6	6

Tableau 2. Les phrases du corpus pour le locuteur entraîné JS.

PB		spontané	acteur	social	mécanique	neutre	réit. séq.	réit. synch.
gynéco	p1	X	X	X	X	X	X	X
venu pour rien	p2	X	X	X	X	X	X	X
saute-mouton	p3	X	X	X	X	X	X	X
total 21		3	3	3	3	3	3	3

Tableau 3. Les phrases du corpus pour le locuteur entraîné PB.

DV		spontané	acteur	social	mécanique	neutre	réit. séq.	réit. synch.
ceinture	d1	X	X	X	X	X	X	X
foot	d2	X	X	X	X	X	X	X
gynéco	d3	X	X	X	X	X	X	X
lise	d4	X	X	X	X	X	X	X
lumière	d5	X	X	X	X	X	X	X
omar	d6	X	X	X	X	X	X	X
waouh	d7	X	X	X	X	X	X	X
passant			X	X	X	X		
mamama foot	m1	X	X	X	X	X	X	X
mamama gynéco		X						X
mamama lumière	m2	X	X	X	X	X	X	X
mamama omar	m3	X	X	X	X	X	X	X
mamama waouh	m4	X	X	X	X	X	X	X
total 83		12	12	12	12	12	11	12

Tableau 4. Les phrases du corpus pour le locuteur naïf DV.

Les phrases issues de la lecture sur écran, comme prévu dans la manipulation de mise en situation, étaient :

ceinture	Il est interdit de frapper sous la ceinture.
foot	Ce qui est énervant dans le football de table, c'est que la ballon tombe souvent de la table.
gynéco	Je connais un gynécologue sourd qui est capable de lire sur les lèvres.
lise	Fais gaffe, Lise, t'as touché une artère !
lumière	Nous vous devons plus que la lumière.
omar	Dis, Omar, t'as pas l'impression qu'on tourne en rond ?
passant	Ce passant chantait beaucoup trop de chansons.
saute-mouton	Les éléphants devraient éviter de jouer à saute-mouton.
waouh	Waouh ! Ça marche ! Essaie là, juste sous mon doigt !

Tableau 5. Les phrases du corpus issues de la tâche prévue de lecture sur écran.

Pour les locuteurs entraînés, qui avaient peu souri pendant les phrases lues, leurs remarques spontanées amusées ont été utilisées aussi.

voir vendredi	Quand il va se voir vendredi !
bonne	Elle est bien bonne.
bonne aussi	Elle est bien bonne aussi.
exagerez	Vous exagerez, hein !
venu pour rien	J'suis venu pour rien, alors ?!

Tableau 6. Les phrases énoncées spontanément par les locuteurs entraînés.

En vue d'une éventuelle récupération des données pour une future synthèse audiovisuelle, les énoncés suivants, représentant trois configurations labiales extrêmes, ont été prononcés :

aJaJaz	C'est pas [aJaJaz] ? (J=fricative post-alvéolaire voisée)
iziziz	C'est pas [iziziz] ?
ibibiz	C'est pas [ibibiz] ? (erreur de prononciation de [iziziz])
ububuz	C'est pas [ybybyz] ?

Tableau 7. Les phrases énoncées pour les configurations labiales extrêmes.

En outre des phrases listées ci-dessus, le corpus contient quelques expressions “ hors condition ” : une répétition en “ ma-ma-ma ” non terminée de “ waouh ” par le locuteur DV ; une énonciation riante de “ gynéco ” produite par PB en dehors d’une des conditions d’enregistrement ; et un rire silencieux (pas de son) pour chaque locuteur, comme référence pour une expression faciale extrême en vue des mesures de l’expression faciale prévues.

8.3. Test de perception

8.3.1. Planification du test de perception

Un test de perception a été préparé pour tester les hypothèses formulées plus haut (voir 0). Pour tester la capacité de distinguer entre différentes expressions, nous avons adopté la méthode de discrimination entre deux stimuli, utilisée par Tartter (1980) et Tartter & Braun (1994). Elle consiste en une présentation des stimuli par paires et de demander lequel des stimuli est plus X.

Comme les mêmes paires de stimuli devaient être présentées en plusieurs conditions, nous avons choisi un ordre de présentation supposé minimiser les effets d’ordre. Les stimuli ont d’abord été présentés de manière unimodale (audio seul, vidéo seule) et ensuite de manière bimodale (audiovisuel). La présentation d’abord en condition audiovisuelle aurait pu influencer sur le jugement en condition audio seul par la suite. Si les stimuli sont présentés d’abord en audio seul, les auditeurs, ne connaissant pas les stimuli, jugent uniquement à partir de ce qu’ils entendent, sans interférence possible avec un souvenir d’images associées au son.

Pour tester les hypothèses 1 à 3, nous avons préparé la présentation des paires suivantes :

- amusement spontané / énonciation neutre en audio seul (23 paires de phrases) ;
- amusement spontané / sourire mécanique à la Tartter, en audio seul (20 paires) ;
- amusement spontané / énonciation neutre en vidéo seule (23 paires) ;
- amusement spontané / énonciation neutre en audiovisuel (23 paires) ;
- amusement spontané / énonciation réitérée en séquentiel, en audiovisuel (23 paires) ;
- amusement spontané / énonciation réitérée en synchrone, en audiovisuel (23 paires).

Une paire consistait donc de deux énonciations de la même phrase, qui se distinguaient par le type d'expression. Dans toutes les parties sauf amusé/mécanique, les paires ont été construites à partir des mêmes 23 phrases produites dans les types d'expression voulus. Les trois phrases (y_1 , y_2 , y_3) n'ayant pas été produites avec un sourire mécanique, elles ne pouvaient pas être utilisées dans la partie "amusé/mécanique", réduisant ainsi le nombre de paires à 20. L'ordre de présentation des deux stimuli dans une paire était randomisé, et les auditeurs devaient choisir laquelle des deux séquences était prononcée par un locuteur plus amusé.

Pour l'hypothèse 4, nous avons rendu la tâche plus difficile. Pour la comparaison des phrases amusées spontanées et des phrases d'acteur amusées, le même protocole qu'avant aurait été peu approprié : d'une part, les expressions d'acteurs étaient exagérées et les signes d'amusement spontané parfois plutôt discret ; d'autre part, l'hypothèse ne porte pas sur le degré d'amusement, mais sur la capacité des juges de distinguer entre parole spontanée et parole produite en tant qu'acteur. Nous avons donc proposé une autre tâche, en demandant d'identifier la manière dont chaque phrase a été produite. Les phrases

- amusées spontanées et
- jouées en tant qu'acteur amusé

ont été présentées une par une, avec la consigne d'identifier si le locuteur avait produit la phrase de manière spontanément amusée ou en tant qu'acteur.

De même, pour tester l'hypothèse 5, nous avons utilisé un test d'identification. Nous avons présentés des stimuli des trois types suivants :

- jouées en tant qu'acteur amusé ;
- jouées en tant qu'acteur avec un sourire de séduction ;
- énoncées avec un sourire mécanique, sans émotion.

Les auditeurs devaient choisir parmi les trois types.

Dans toutes les conditions, les juges étaient forcés de choisir, avec la possibilité de cocher une petite case "pas sûr" s'ils étaient très incertains de leur choix. Nous avons ainsi adopté la même position que Tartter & Braun (1994) qui ont forcé leurs juges à deviner parce qu'ils supposaient que les juges savaient plus que ce qu'ils ne pensaient.

8.3.2. Un programme Supercard pour le test

Nous avons écrit un programme Supercard pour présenter de manière interactive les stimuli et pour enregistrer les choix des juges.

Pour pouvoir utiliser le programme sur deux ordinateurs différents en même temps, tous les accès du disque dur étaient faits relatifs à un chemin d'accès principal choisi par l'utilisateur au démarrage du programme. Les informations sur la présentation des stimuli, c'est-à-dire

ordre des paires, noms des fichiers constituant une paire, et ordre des deux fichiers dans une paire, ont été préparé dans des fichiers texte individuels, un fichier par condition de présentation. Ceci garantissait une modularité maximale du programme, permettant des changements sans beaucoup de difficultés.

Les choix de l'utilisateur étaient enregistrés dans des variables au fur et à mesure du déroulement du programme, et écrit sur disque dur à la fin, dans un fichier intitulé par nom, âge, et sexe de l'utilisateur, et dans un format facilitant l'analyse statistique ensuite. Pour chaque choix, quatre informations étaient enregistrées, séparées par des signes de tabulation :

- un point d'interrogation était marqué en début de ligne si la case " pas sûr " était cochée, sinon le premier champ restait vide ;
- un chiffre 1, si la réponse était " juste " (comme attendue), un 0 sinon ;
- un mot pour identifier le choix fait (" amuse ", " neutre ", " mecano ", " re_seq ", " re_syn ", pour les choix de discrimination ; " naturel ", " acteur ", " mecano ", " seduc " pour les choix d'identification ;
- et le nom du fichier choisi pour identifier le stimulus.

8.3.3. *Le protocole du test de perception*

Les tests ont été menés avec 20 juges naïfs de 19 à 50 ans (11 femmes, 9 hommes) qui ont passé les tests individuellement, dans une salle informatique silencieuse. Les stimuli étaient présentés par écouteurs et sur écran 17". Chaque session durait 45 minutes. Pour habituer les juges à la tâche demandée, trois paires de stimuli audiovisuels¹⁷ étaient proposées avant le début du test.

8.3.4. *Commentaires des juges*

L'un des juges (AA50m), spécialiste de la prise du son, a indiqué qu'il pouvait " tricher " en condition audio seul, à cause du bruit du ventilateur du Macintosh qui avait été enregistré différemment selon la prise de son (emplacement du micro) dans les différentes conditions. Comme ses taux de reconnaissance étaient en général comparables à ceux des autres juges (à l'exception de la condition amusé-sourire mécanique audio seul, où il avait des scores nettement supérieurs aux autres), ses résultats n'ont pas été enlevées de l'analyse.

Plusieurs juges se sont demandés si nous n'avions pas parfois proposé deux stimuli audiovisuels identiques dans une paire (ce qui n'est pas le cas).

La plupart des juges trouvaient la dernière partie (test d'identification parmi acteur/séduction/sourire mécanique) très difficile. Certains suggéraient des différences possibles entre juges hommes et femmes quant à la perception de la séduction. Une juge

¹⁷Pour éviter de présenter dans le pré-test des paires de stimuli faisant partie du test, nous avons proposé dans le pré-test des oppositions de stimuli non contenues dans le test, telles que acteur/neutre.

mentionnait la contribution à l'expression de la séduction pour le locuteur DV d'un petit "click" produit avec la langue avant la phrase. Ce serait un indice local à influence sur la perception globale de l'énoncé.

8.4. Les résultats du test de perception

Les résultats du test de perception ont été analysés en partie sous Excel, en partie par un test de signification ANOVA¹⁸.

8.4.1. L'opposition amusé–neutre

Pour les 23 paires de phrases amusé–neutre présentées dans les trois conditions – audio seul, vidéo seule, et audiovisuel –, les taux d'identification des stimuli "amusé" sont de 84,3% (audio), 95,2% (vidéo), et 93,5% (audiovisuel). La différence entre le taux moyen en condition audio d'un côté et en conditions vidéo et audiovisuelle de l'autre côté est significative ($p < 0.01$, test de Tukey). La différence entre les taux en conditions vidéo et audiovisuelle n'est pas significative.

Ces résultats montrent deux choses. Tout d'abord, en audio seul, les stimuli "amusé" et "neutre" sont perceptivement distingués, comme le laissent supposer les résultats de Tartter (1980) et de Tartter & Braun (1994) : ils avaient présenté des paires d'énoncés produits avec/sans sourire mécanique en audio seul et obtenu des taux d'identification de 63,4% (Tartter, 1980 : logatomes et phrases pleines, locuteurs non mélangés dans le test de perception) et de 54,4% (Tartter & Braun, 1994 : seulement logatomes, locuteurs mélangés dans le test de perception).

Ensuite, le fait qu'en condition vidéo, les stimuli "amusé" aient été bien reconnus, prouve la présence d'indices visuels pour l'amusement. Les indices acoustiques et visuels pour l'amusement peuvent fonctionner séparément. La présentation des stimuli en condition audiovisuelle n'augmente pas le taux d'identification au-dessus du niveau de la condition vidéo seule. L'expression et la perception de l'amusement se font donc effectivement selon les deux canaux audio et visuel, conformément à notre hypothèse (1).

La Figure 4 présente de manière graphique les taux moyens d'identification correcte à travers les 20 juges, détaillés selon phrase et condition.

¹⁸ Je remercie Marie Cathiard de son aide gentiment proposée, et qui a rendu possible les analyses ANOVA..

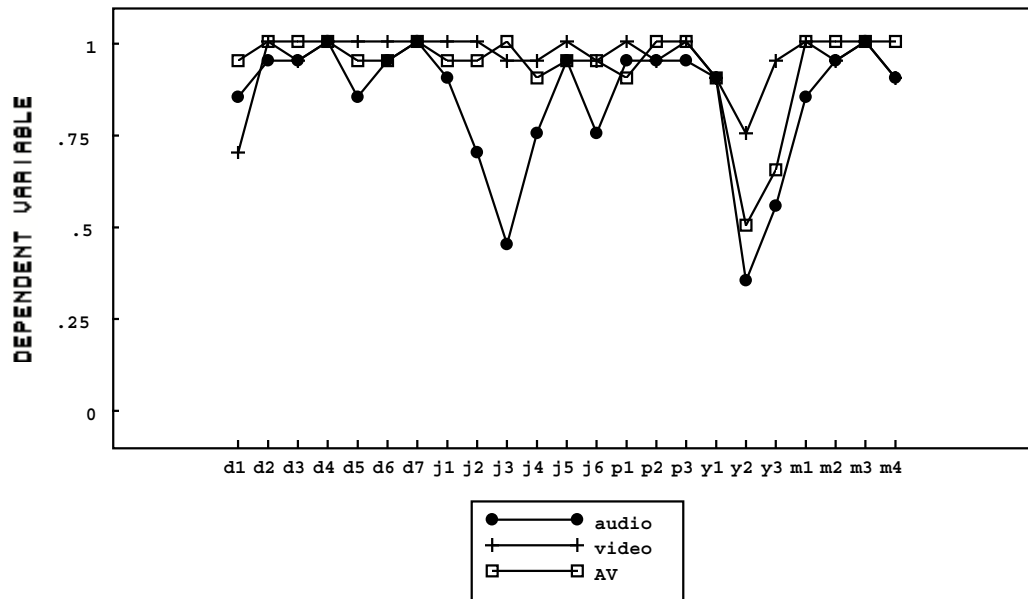


Figure 4. Les taux d'identification correcte pour les 23 paires de phrases amusé-neutre.

A l'exception de la phrase (d1), le taux d'identification en condition vidéo est toujours supérieur à celui en condition audio. De l'autre côté, quand le taux est élevé en audio, il l'est aussi dans les deux autres conditions.

L'analyse détaillant les phrases permet d'identifier des expressions d'amusement mal reconnues. D'une part, pour deux phrases (y2, y3), les deux types de stimuli sont mal distingués : les taux d'identification en condition audiovisuelle étant près du hasard (50%). D'autre part, quatre phrases sur six du locuteur JS (j2, j3, j4, j6) ont été bien reconnues en vidéo (95% à 100%) et en audiovisuel (90% à 100%), mais mal identifiées par l'audio seul (45% pour j3, de 70% à 75% pour j2, j4, j6).

L'analyse ci-dessus ne rend pas directement compte du facteur locuteur. Pour ce faire, nous avons calculé, pour chaque locuteur, le taux moyen de reconnaissance des phrases pleines énoncées par ce locuteur (les phrases d1 à d7 pour le locuteur DV ; j1 à j6 pour JS ; p1 à p3 pour PB ; y1 à y3 pour YM). Ainsi, il est possible de chercher des effets de locuteur dans une condition de présentation particulière ou à travers les trois conditions (Figure 5).

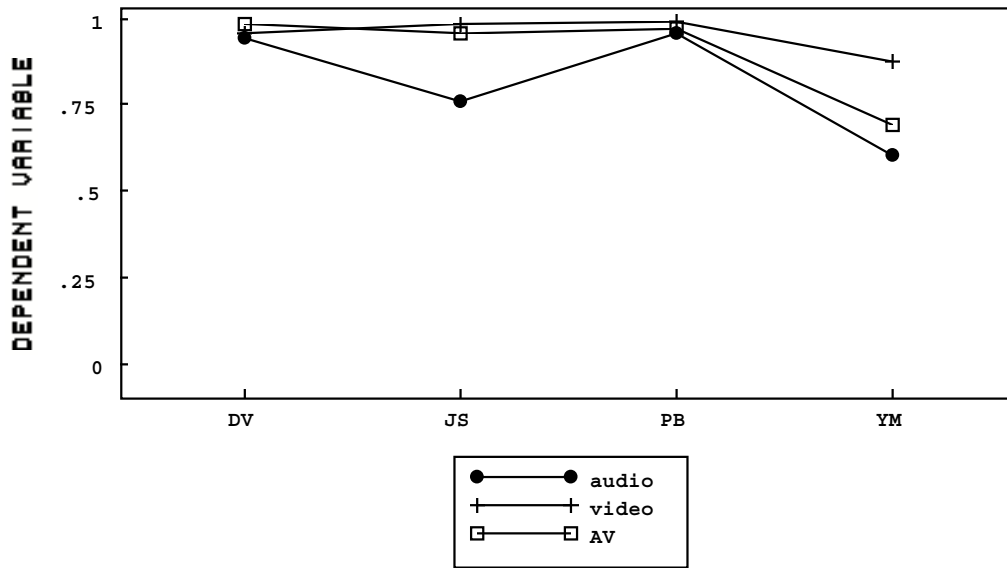


Figure 5. Les taux d'identification par locuteur pour les paires de phrases amusé-neutre dans les trois conditions de présentation (audio, vidéo, audiovisuel).

Les taux globaux d'identification, moyennés à travers les conditions de présentation, sont de 95,5% (DV), 89,2% (JS), 96,7% (PB), et 71,7% (YM). Le taux d'identification pour le locuteur YM est inférieur à celui de chaque autre locuteurs ($p < .01$, test de Tukey). Le taux pour JS est inférieur à celui pour PB ($p < .05$, test de Tukey).

Les taux d'identification par condition de présentation sont de 80,9% (audio), 94,3% (vidéo), et 89,5% (AV). Ces valeurs ne sont évidemment pas les mêmes que pour l'ensemble des phrases : d'une part, les quatre phrases en "ma-ma-ma" n'entrent pas dans l'analyse ; d'autre part, les moyennes par locuteur ne sont pas pondérées selon le nombre de phrases, pour égaliser les poids avec lesquels les locuteurs contribuent à la moyenne. Néanmoins, les différences entre conditions sont les mêmes que dans l'analyse détaillée par phrases : le taux d'identification en audio seul est inférieur à ceux en vidéo seule et en audiovisuel ($p < .01$, test de Tukey), tandis que la différence entre la condition vidéo et audiovisuel n'est pas significative.

La comparaison par paires (test de Tukey) des taux par locuteur et par condition indique essentiellement que les taux d'identification JS/audio, YM/audio et YM/audiovisuel sont inférieurs au reste de manière significative (détails voir Annexe).

La mauvaise distinction entre les phrases "amusé" et "neutre" pour le locuteur YM peut être due au fait que, faute d'énonciation neutre, nous avons classé la deuxième énonciation d'une phrase *dans la mise en situation* comme phrase "neutre". Les mauvais scores de distinction indiquent donc que le locuteur n'exprimait pas beaucoup plus d'amusement pendant la première énonciation que pendant la deuxième.

8.4.2. Les effets prosodiques de l'amusement

Pour les 20 paires amusé–sourire mécanique présentées en condition audio seul, le taux d'identification des phrases “amusé” est de 68,5%. La capacité des juges à reconnaître l'amusement spontané comme exprimant plus d'amusement que le sourire mécanique semble indiquer la présence d'indices prosodiques. En condition amusé–neutre, audio seul, le taux d'identification pour les mêmes 20 phrases est de 88,0%¹⁹. La différence entre ces deux conditions est significative ($p < 0,01$, test de Tukey). Comme on pourrait s'y attendre, il est donc plus difficile de distinguer une expression d'amusement d'un sourire mécanique que de la distinguer d'une énonciation neutre.

La présentation graphique des taux d'identification par phrase (Figure 6) suggère qu'il faut distinguer pour l'opposition amusé–mécanique deux groupes de phrases :

- pour les phrases (d1, d2, d6, j1, j2, j3, j4, p3), les juges semblent avoir répondu au hasard, il n'y a donc pas d'indices permettant de distinguer entre amusement spontané et sourire mécanique ;
- pour les phrases (d3, d4, d5, d7, j5, j6, p1, p2, m1, m2, m3, m4), les juges ont réussi l'identification presque aussi bien que pour l'opposition amusé–neutre. Pour une analyse acoustique permettant de dégager des indices prosodiques de l'amusement, ces phrases, dans les trois conditions de production “amusé”, “sourire mécanique”, et “neutre”, semblent donc se prêter particulièrement.

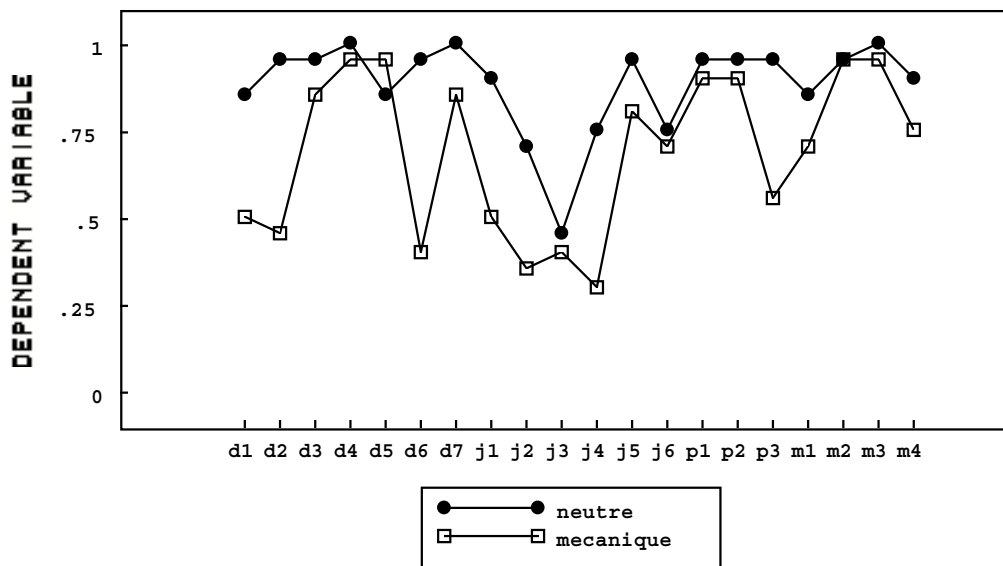


Figure 6. Les taux d'identification correcte pour les 20 paires de phrases amusé–neutre et amusé–sourire mécanique présentées en audio seul.

¹⁹ Le taux moyen d'identification amusé–neutre en condition audio est de 84,3% moyenné à travers les 23 phrases du test, et de 88,0% moyenné à travers les 20 phrases qui existent aussi en amusé–mécanique. Pour la comparaison entre les deux conditions, il fallait enlever (y1, y2, y3) qui n'ont pas d'équivalent produit avec un sourire mécanique.

8.4.3. Simulation par répétition

Les résultats des oppositions entre l'amusement spontané et les deux types de répétition sont plus difficiles à interpréter. Le taux moyen²⁰ pour les 23 paires amusé-répétition séquentielle est de 55,9%, pour les 23 paires amusé-répétition synchrone de 67,4%. Une incapacité des juges de distinguer entre l'original et sa répétition indiquerait une simulation bien réussie. Bien que le taux moyen à travers phrases de 55,9% suggère que c'est le cas pour la répétition séquentielle, la présentation des détails par phrase montre une situation plus complexe :

- pour les phrases (d4, p2, p3, y1, m1, m2, m3, m4), les originaux ont été reconnus comme plus "amusé" dans plus de 75% des cas ;
- par contre, pour les phrases (d1, d6, j2, j4, j6), c'est la phrase répétée qui est jugée plus "amusé" que l'original dans plus de 75% des cas ;
- les phrases (d2, d3, d5, d7, j1, j3, j5, p1, y2, y3) sont reconnues plus ou moins au hasard.

La situation pour la répétition synchrone est similaire, sauf que les répétitions ont, en moyenne à travers juges, rarement été jugées plus "amusé" que les originaux :

- les originaux des phrases (d2, d3, d4, d5, d7, p1, p2, y1, y2, y3, m3, m4) ont été identifiés dans plus de 75% des jugements, indiquant éventuellement une mauvaise qualité des répétitions ;
- pour les phrases (d1, j1, j2), la répétition a été jugée plus "amusé" que l'original ;
- et pour les phrases (d6, j3, j4, j5, j6, p3, m1, m2), les choix sont plus ou moins au hasard, ce qui pourrait indiquer une simulation convaincante.

²⁰ Taux moyen d'identification à travers locuteurs et à travers phrases pour la phrase produite de manière spontanément amusée comme exprimant plus d'amusement que la phrase répétée.

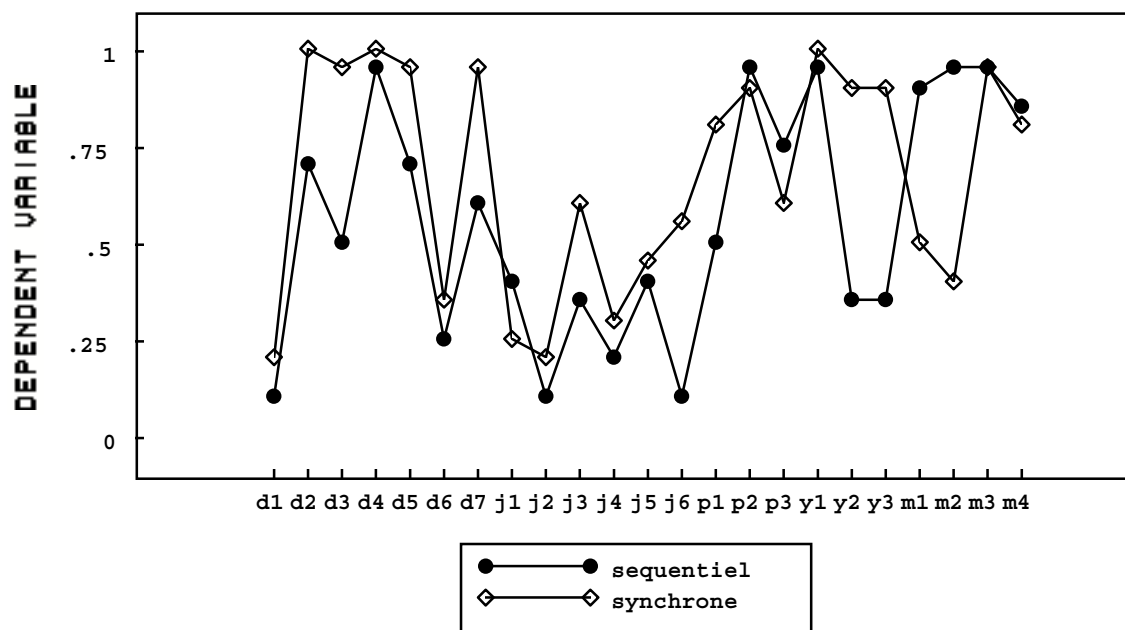


Figure 7. Les taux d'identification correcte pour les 23 paires de phrases amusé-réitéré en séquentiel et amusé-réitéré en synchrone présentées en audiovisuel.

8.4.4. Spontané et acteur

Pour la tâche d'identification du mode de production d'un énoncé (amusement spontané / acteur jouant d'être amusé), le taux moyen d'identification correcte est de 58,8%. A cet égard, il n'y a pas de différence entre l'identification des stimuli spontanés (58,3%) et des stimuli acteurs (59,2%). Ce qui est remarquable, par contre, est la variabilité entre juges : parmi les 20 juges,

- 8 juges (6 femmes, 2 hommes) avaient des taux d'identification correcte entre 66,7% et 89,6%, ils étaient donc plutôt capable d'identifier la nature du stimulus ;
- et 12 juges avaient des taux d'identification correcte entre 33,3% et 58,3%, c'est-à-dire qu'ils ont répondu plus ou moins au hasard.

D'autre part, il faut de nouveau remarquer une grande variabilité inter-phrased :

- certaines phrases (spontané : d4, d5, j3, j5, p1, y4, m1, m2, m3 ; acteur : d2, d5, d6, p3, y1, y2, y3, y4, m2) ont été reconnues correctement dans au moins 70% des cas ;
- quelques phrases (production spontanée : d2, j4, m4 ; acteur : j6, p2) ont été mal identifiées dans au moins 70% des cas ;

- les autres phrases (spontané : d1, d3, d6, d7, j1, j2, j6, p2, p3, y1, y2, y3 ; acteur : d1, d3, d4, d7, j1, j2, j3, j4, j5, p1, m1, m3, m4) ont été reconnues plus ou moins au hasard.

Quelles sont les raisons de ces variations importantes entre juges et entre phrases ? En ce qui concerne la variabilité inter-phrased, une explication pourrait être que notre protocole expérimental pour l'enregistrement de parole exprimant de l'amusement spontané ne mènerait pas à une énonciation suffisamment spontanée. Cependant, même parmi les phrases produites spontanément par les locuteurs professionnels (d1, d2, d3, p2, y4), seules deux (j3, y4) sont bien reconnues comme spontanées. Cette explication ne rend pas non plus compte de la mauvaise identification des phrases "acteur".

D'autre part, si on rejette l'hypothèse d'une distinction perceptive entre l'amusement spontané et joué, il est difficile d'interpréter les bons scores obtenus par 8 des 20 juges. Peut-être faut-il interpréter les résultats comme indiquant une capacité particulière des juges qui réussissent bien à cette tâche ? Dans ce cas, il serait important de développer une méthode pour identifier cette capacité.

En ce qui concerne le rôle hypothétique du plissement de la peau autour des yeux pour l'identification de l'amusement spontané (Ekman et al., 1990 ; Damasio, 1994), des mesures vidéo des enregistrements bien reconnus comme étant spontané ou acteur pourraient être effectuées et comparées.

8.4.5. Acteur d'amusement, de séduction, et sourire mécanique

Il est très intéressant de noter que les résultats sont meilleurs dans la partie ressentie comme difficile par la majorité des juges. Le taux moyen d'identification pour tous les stimuli est de 54,0%, comparé à un niveau de hasard de 33,3%. Les confusions entre catégories sont indiquées dans le Tableau 8 pour l'ensemble des juges. Quelques juges ont suggéré des différences possibles entre juges masculins et féminins quant à l'interprétation des stimuli de séduction. Comme il s'agit d'un signal social dirigé vers des individus du sexe opposé, et comme tous nos locuteurs sont des hommes nous avons calculé les confusions séparément, dans le Tableau 9 pour les juges masculins, et dans le Tableau 10 pour les juges féminins. Les principales tendances de la matrice de confusion pour tous les juges sont présentées de manière graphique dans la Figure 8.

jugements	stimuli			
	acteur	séduction	sourire mécanique	tous stimuli confondus
acteur	63,5%	18,3%	17,9%	34,2%
séduction	12,7%	41,2%	22,4%	25,6%
sourire mécanique	23,8%	40,6%	59,8%	40,1%
total	100,0%	100,0%	100,0%	100,0%

Tableau 8. La matrice de confusion pour le test d'identification acteur-séduction-sourire mécanique pour l'ensemble des juges.

jugements	Stimuli			
	acteur	séduction	sourire mécanique	tous stimuli confondus
acteur	60,7%	21,4%	18,5%	34,6%
séduction	12,0%	40,2%	27,0%	26,3%
sourire mécanique	27,4%	38,5%	54,5%	39,1%
total	100,0%	100,0%	100,0%	100,0%

Tableau 9. La matrice de confusion pour le test d'identification acteur-séduction-sourire mécanique pour les juges masculins.

jugements	Stimuli			
	acteur	séduction	sourire mécanique	tous stimuli confondus
acteur	65,7%	15,7%	17,3%	34,0%
séduction	13,3%	42,0%	18,6%	25,0%
sourire mécanique	21,0%	42,3%	64,1%	41,0%
total	100,0%	100,0%	100,0%	100,0%

Tableau 10. La matrice de confusion pour le test d'identification acteur-séduction-sourire mécanique pour les juges féminins.

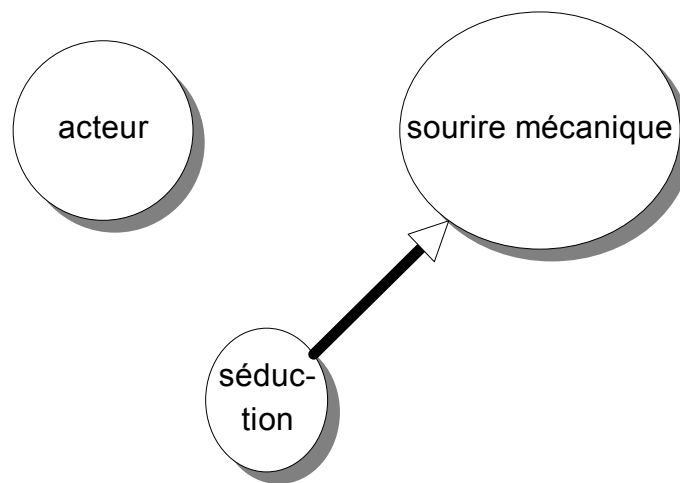


Figure 8. Présentation graphique des principales tendances de la matrice de confusion pour tous les juges. Une flèche indique qu'un stimulus de la catégorie d'où part la flèche a été jugé comme étant la catégorie sur laquelle aboutit la flèche en plus de 33% des cas (= hasard). La taille des points représente l'attraction pour les jugements d'une catégorie : plus un point est petit, moins la catégorie correspondante a été choisie dans le test de perception.

Les tendances suivantes sont visibles dans les matrices de confusion, de manière comparable pour les juges des deux sexes. Avant tout, les stimuli de séduction ont été interprétés comme sourires mécaniques aussi souvent qu'ils ont été reconnus comme séduction. Le phénomène n'est pas symétrique, c'est-à-dire que les sourires mécaniques ont été bien reconnus. Par conséquent, la catégorie "sourire mécanique" a été choisie plus souvent que la moyenne, et la catégorie "séduction" moins souvent²¹.

Cependant, les stimuli produits avec une intention d'exprimer de la séduction ne sont pas un ensemble homogène :

- toutes les phrases pleines du locuteur DV, une phrase répétée en "ma-ma-ma", une des phrases du locuteur JS et une des phrases du locuteur YM, produites avec une intention d'exprimer de la séduction, sont reconnues comme séduction dans 55% à 80% des cas ;
- les autres phrases produites comme séduction ont des taux d'identification au-dessous de 40%. Ceci est particulièrement vrai pour le locuteur PB, pour qui les phrases produites avec l'intention d'exprimer de la séduction sont reconnues de manière fiable comme des sourires mécaniques (75% à 80%). Pour ces phrases, nous supposons que les locuteurs n'ont pas réussi à exprimer de la séduction.

²¹ Ces observations se basent sur la dernière colonne dans les matrices de confusions. Elle est calculé par le nombre total de choix pour une catégorie, indépendamment du type de stimulus, relatif au nombre total de choix dans toutes les catégories.

Ces résultats montrent la nécessité de disposer d'un moyen d'évaluation de la qualité de l'expression, un critère pour évaluer les stimuli avant de les présenter dans un test de perception. La solution de Banse & Scherer (1996) et de Leinonen (1997) est de présélectionner les stimuli par des jugements d'experts (voir 0).

CONCLUSIONS ET PERSPECTIVES

L'établissement d'un corpus audiovisuel de parole (contenant, sur les mêmes phrases, des expressions de l'amusement spontané et joué (acteur), des simulations de l'amusement spontané par répétition, un sourire social "de séduction", un sourire mécanique, et une énonciation sans émotion), et la validation de ce corpus par un test de perception, sont les premiers pas vers une synthèse audiovisuelle de l'amusement spontané.

Les résultats du test de perception confirment l'hypothèse de l'expression bimodale (audiovisuel) de l'amusement spontané pendant une tâche de parole (en français), et de la capacité de récupération perceptive des juges autant à partir d'un seul canal (audio seul / vidéo seule) qu'à partir de leur combinaison (présentation audiovisuelle). Le taux de discrimination entre une expression amusée spontanée et une expression sans émotion des mêmes énoncés est inférieur en audio seul (84%) qu'en vidéo seule et audiovisuel ($\approx 94\%$), mais clairement différente du hasard. Le taux en condition audiovisuelle n'est pas supérieur au taux en condition vidéo seule.

Pour l'opposition amusé-neutre, nous avons observé un effet du locuteur : les phrases "amusé" et "neutre" sont significativement moins bien distinguées pour l'un des locuteurs (YM) que pour les autres. Cette différence peut être due au fait que, pour ce locuteur, l'énonciation "neutre" n'a pas été produite dans les mêmes conditions que pour les autres locuteurs. Pour un autre locuteur (JS), le taux d'identification est significativement inférieur aux taux des "meilleurs" locuteurs (DV, PB) en condition de présentation audio seul.

Ensuite, la capacité des juges à distinguer entre l'expression spontanée de l'amusement et un sourire mécanique (taux de discrimination : 68,5%) montre que l'amusement spontané s'exprime dans la voix par des indices autres que les simples conséquences acoustiques du geste du sourire.

Les résultats des tests opposant un original spontané à sa répétition en séquentiel ou en synchrone sont complexes à interpréter, du fait d'une grande variation inter-phrases : pour certaines phrases, l'original est jugé plus amusé que sa répétition, pour d'autres c'est le contraire, et pour un troisième groupe de phrases, les réponses semblent suivre le hasard. Il est néanmoins possible de tirer les conclusions suivantes : pour les deux types de répétition, l'original semble reconnaissable (dans 55,9% des cas en moyenne pour l'opposition amusé-répétition séquentielle, et dans 67,4% des cas pour l'opposition amusé-répétition synchrone) ; la répétition séquentielle, plus confondue avec l'original, semble être une meilleure simulation que la répétition synchrone, relativement bien distincte de l'original.

Pour la tâche d'identification des phrases spontanées/acteur, huit juges sur vingt avaient des bons taux d'identification (au-dessus de 66,7%). Une interprétation possible de ce résultat est d'attribuer une capacité particulière à ces juges.

La tâche d'identification parmi des sourires d'acteur d'amusement, acteur de séduction et sourire mécanique a été bien réussie (en moyenne 54% d'identification correcte comparé à un niveau du hasard de 33,3%). Les "sourires de séduction" ont été moins bien reconnus en moyenne que les deux autres types de sourire. Cependant, il faut distinguer une classe d'énoncés de "séduction" bien reconnus d'une classe d'énoncés de "séduction" mal reconnus. Nous supposons que pour cette dernière, les locuteurs n'ont pas réussi à exprimer la séduction. Ainsi, autant l'expression de l'amusement que l'expression de la séduction comportent des indices spécifiques permettant leur identification.

La suite logique du présent travail est une analyse vidéo de l'expression faciale et une analyse audio du signal acoustique. L'analyse vidéo pourra obtenir des indices objectifs pour la présence d'un sourire, et vérifier une éventuelle influence du plissement de la peau autour des yeux sur les jugements spontané/acteur. Par l'analyse acoustique, on cherchera notamment des différences prosodiques systématiques entre l'amusement spontané et le sourire mécanique, en se basant sur les paires bien distinguées en condition audio seul dans le test de perception.

BIBLIOGRAPHIE

1. Articles liés aux émotions

a) Les articles que nous avons lu

- ALT, S. (1997) : Prosodie attitudinale de l'allemand : Indices sur la globalité des réalisations intonatives de quelques particules illocutoires, *mémoire de DEA en Sciences du Langage*, Université Stendhal-Grenoble 3.
- AUBERGÉ, V., GRÉPILLAT, T., & RILLIARD, A. (1997) : Can we perceive attitudes before the end of sentences ? The gating paradigm for prosodic contours, *Eurospeech 1997*, Athen.
- BANSE, R., & SCHERER, K. R. (1996) : Acoustic Profiles in Vocal Emotion Expression, *Journal of Personality and Social Psychology*, 170 (3), p. 614-636.
- DAMASIO, A. R. (1994) : *Descartes' Error. Emotion, Reason, and the Human Brain*. A. Grosset / Putnam Books. (traduction française (1995) : *L'erreur de Descartes. La raison des émotions*. Paris : Odile Jacob.)
- EKMAN, P., DAVIDSON, R. J., & FRIESEN, W. V. (1990) : Duchenne's smile: Emotional expression and brain physiology II, *J. Pers. Soc. Psych.*, 58, p. 342-353.
- FRIED, I., WILSON, C. L., MACDONALD, K. A., & BEHNKE, E. J. (1998) : Electric current stimulates laughter, *Nature*, 391 (12 february 1998), p. 650.
- GERRARDS-HESSE, A., SPIES, K., & HESSE, F. W. (1994) : Experimental inductions of emotional states and their effectiveness: A review, *British Journal of Psychology*, 85, p. 55-78.
- JOHNSTONE, T., BANSE, R., & SCHERER, K. R. (1995) : Acoustic profiles in prototypical vocal expressions of emotion, *Proceedings of the 13th ICPhS, Stockholm*, Vol. 4, p. 2-5.
- LADD, D. R., SILVERMAN, K. E. A., TOLKMITT, F., BERGMANN, G., & SCHERER, K. R. (1985) : Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect, *Journal of the Acoustic Society of America*, 78 (2), p. 435-444.
- LEINONEN, L., HILTUNEN, T., LINNANKOSKI, I., & LAAKSO, M. (1997) : Expression of emotional-motivational connotations with a one-word utterance, *Journal of the Acoustic Society of America*, 102 (3), p. 1853-1863.
- MORONI, V. (1997) : Enquête sur les attitudes du français : définition et interprétation, *mémoire de maîtrise en Sciences du Langage, mention Industries de la Langue*, Université Stendhal-Grenoble 3.
- MURRAY, I. R., & ARNOTT, J. L. (1993) : Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, *Journal of the Acoustic Society of America*, 93 (2), p. 1097-1108.
- OHALA, J. J. (1996) : Ethological Theorie and the expression of emotion in the voice, *ICSLP 96*.
- PLOOG, D. (1979) : Phonation, emotion, cognition, with reference to the brain mechanisms involved, *Brain and Mind, CIBA Foundation Symposium*, p. 79-98. Amsterdam : Elsevier/North-Holland.
- PROVINE, R. R. (1996) : Laughter, *American Scientist*, 84 (january-february), p. 38-45.
- SCHAUB, H. (1995) : Die Rolle der Emotionen bei der Modellierung kognitiver Prozesse, *Workshop Artificial Life 12.-13.10.1995, GMD-Sankt Augustin*,
<http://www.uni-bamberg.de/~ba2dp1/private/schaub/papers/al95.htm>.
- SCHERER, K. R. (1986) : Vocal Affect Expression : A Review and a Model for Future Research, *Psychological Bulletin*, 99, p. 143-165.
- SCHERER, K. R. (1996) : Adding the Affective Dimension: A new look in Speech Analysis and Synthesis, *ICSLP 96*.
- SCHERER, K. R., LADD, D. R., & SILVERMAN, K. E. A. (1984) : Vocal cues to speaker affect: Testing two models, *Journal of the Acoustic Society of America*, 76 (5), p. 1346-1356.
- SCHERER, K. R. (1989) : Les émotions : fonctions et composantes, in *Les émotions* (B. Rimé & K. R. Scherer, eds.), p. 87-133. Neuchâtel ; Paris : Delachaux-Niestlé.

- TARTTER, V. C. (1980) : Happy talk: Perceptual and acoustic affects of smiling on speech, *Perception & Psychophysics*, 27 (1), p. 24-27.
- TARTTER, V. C., & BRAUN, D. (1994) : Hearing smiles and frowns in normal and whisper registers, *J. Acoust. Soc. Am.*, 96 (4), p. 2101-2107.

b) Articles intéressants, que l'on pourrait lire

- ARNDT, H., & JANNEY, R. W. (1991) : Verbal, prosodic, and kinesic emotive contrasts in speech, *Journal of Pragmatics*, 15, p. 521-549.
- ARNOLD, M. B. (1960) : *Emotion and personality, Vol. 1, Psychological Aspects*. New York : Columbia University Press.
- AVERILL, J. R. (1975) : A semantic atlas of emotional concepts, *JSAS Catalogue of Selected Documents in Psychology*, 5, p. 330.
- BAUER, H. R. (1987) : The frequency code: oral-facial correlates of fundamental frequency, *Phonetica*, 44, p. 173-191.
- BESKOW, J. (1995) : Rule-based visual speech synthesis, *Eurospeech 1995, Madrid*, p. 299-302.
- BEZOOIJEN, R. VAN (1984) : *The characteristics and recognizability of vocal expression of emotions*. Dordrecht, The Netherlands: Foris.
- BOROD, J. C. (1992) : Interhemispheric and intrahemispheric control of emotion : A focus on unilateral brain damage, *Journal of Consulting and Clinical Psychology*, 60, p. 339-348.
- BUCK, R. (1984) : *The Communication of Emotion*. New York : Guildford.
- CACIOPPO, J. T., KLEIN, D. J., BERNTSON, G. C., & HATFIELD, E. (1993) : The psychophysiology of emotion, in *Handbook of emotions* (M. Lewis & J. M. Haviland, eds.), p. 119-142. New York: Guilford Press.
- CAHN, J. E. (1988) : From Sad to Glad: Emotional Computer Voices, *Proceedings of Speech Tech '88, New York*, p. 35-36.
- CAHN, J. E. (1989) : Generation of Affect in Synthesized Speech, *Proceedings of AVIOS '89, Meeting of the American Voice Input/Output Society*.
- CEFFL, C. & JANNEY, R. W. (1994) : Toward a pragmatics of emotive communication, *Journal of Pragmatics*, 22, p. 325-373.
- CHRISTIANSON, S. (1992) : *The handbook of emotions and memory: research and theory*. Hillsdale, N.J. ; Hove, London: L. Erlbaum.
- DAVIDSON, R. (1992) : Prolegomenon to emotion : Gleanings from Neuropsychology, *Cognition and Emotion*, 6, p. 245-268.
- DAVIDSON, R. J., EKMAN, P., SARON, C., SENULIS, J., & FRIESEN, W. V. (1990) : Emotional expression and brain physiology I: Approach / withdrawal and cerebral asymmetry, *J. Pers. Soc. Psych.*, 58, p. 330-341.
- DAVITZ, J. R. (1964a) : A Review of Research Concerned with Facial and Vocal Expressions of Emotion, in *The Communication of Emotional Meaning* (J. R. Davitz, ed.), p. 13-30. New York : McGraw-Hill.
- DAVITZ, J. R. (1964b) : Personality, Perceptual, and Cognitive Correlates of Emotional Sensitivity, in *The Communication of Emotional Meaning* (J. R. Davitz, ed.), p. 57-68. New York : McGraw-Hill.
- DAVITZ, J. R. (1964c) : Auditory Correlates of Vocal Expressions of Emotional Feeling, in *The Communication of Emotional Meaning* (J. R. Davitz, ed.), p. 101-112. New York : McGraw-Hill.
- DAVITZ, J. R., & DAVITZ, L. J. (1959a) : The Communication of Feelings by Content-Free Speech, *J. Commun.*, 9, p. 6-13.
- DAVITZ, J. R., & DAVITZ, L. J. (1959b) : Correlates of Accuracy in the Communication of Feelings, *J. Commun.*, 9, p. 110-117.
- DÖRNER, D., SCHAUB, H., STÄUDEL, T., & STROHSCHNEIDER, S. (1988) : Ein System zur Handlungsregulation oder: Die Interaktion von Emotion, Kognition und Motivation, *Sprache & Kognition*, 4, p. 217-232.
- EKMAN, P., & DAVIDSON, R. J. (1993) : Voluntary smiling changes regional brain activity, *Psychological Science*, 4, p. 342-345.
- EKMAN, P., & FRIESEN, W. V. (1969) : The Repertoire of Nonverbal Behavior: Categories, Origins, Usage and Coding, *Semiotica*, 1, p. 49-98.

- EKMAN, P., & FRIESEN, W. V. (1976) : Measuring facial movement. *Journal of Environmental Psychology and Nonverbal Behavior*, 1, 56-75.
- EKMAN, P., & FRIESEN, W. V. (1978) : *The facial action coding system*. Palo Alto, California : Consulting Psychologists Press.
- FRICK, R. W. (1985) : Communicating Emotion: The Rôle of Prosodic Features, *Psychological Bulletin*, 97, p. 412-429.
- FRICK, R. W. (1986) : The Prosodic Expression of Anger: Differentiating Threat and Frustration, *Aggress. Behav.*, 12, p. 121-128.
- GAINOTTI, G. (1972) : Emotional behavior and hemispheric side of the lesion, *Cortex*, 8, p.41-55.
- GANDOUR, J., et al. (1995) : Speech prosody in Affective Contexts in Thai Patients with Right Hemisphere Lesions, *Brain. Lang.*, 51 (3), p. 422-443.
- GARDNER, H., BROWNELL, H., WAPNER, W., & MICHELOW, D. (1983) : Missing the point : The role of the right hemisphere in the processing of complex linguistic materials, in *Cognitive Processes and the Right Hemisphere* (E. Pericman, ed.). New York : Academic Press.
- HAUSER, M. D. (1996) : *The evolution of communication*. Cambridge, Massachusetts ; London, England : MIT Press.
- HEILMAN, K., WATSON, R. T., & BOWERS, D. (1983) : Affective disorders associated with hemispheric disease, in *Neuropsychology of Human Emotion* (K. Heilman & P. Sath, eds.), p.45-64. New York : Guilford Press.
- IZARD, C. E. (1977a) : *Human Emotions*. New York : Plenum.
- IZARD, C. E. (1977b) : *A social interactional theory of emotions*. New York : Wiley.
- KAPPAS, A., HESS, U., & SCHERER, K. R. (1991) : Voice and emotion, in *Fundamentals of nonverbal behavior* (R. S. Feldman, B. Rimé, eds.), p. 200-238. Cambridge, England: Cambridge UP.
- KITAYAMA, SH., & MARKUS, H. R. (1994) : *Emotion and Culture. Empirical Studies of Mutual Influences*. ISBN: 1-55798-224-4.
- LAZARUS, R. S. (1966) : *Psychological stress and the coping process*. New York : McGraw Hill.
- LEVENTHAL, H. (1980) : Toward a comprehensive theory of emotion, in *Advances in Social Psychology*, vol 13 (L. Berkowitz, ed.). New York : Academic Press.
- LEVI, L. (1975) : *Emotions: their parameters and measurement*.
- MANDLER, G. (1975) : *Mind and emotions*. New York : Wiley.
- MURRAY, I. R. (1989) : Simulating Emotion in Synthetic Speech, *Ph. D. thesis*, University of Dundee, UK.
- MURRAY, I. R., ARNOTT, J. L., & NEWELL, A. F. (1988) : HAMLET – Simulating Emotion in Synthetic Speech, *Proceedings of Speech '88, The 7th FASE Symposium, Edinburgh*, Vol. 4, p. 1217-1223.
- NIEMEIER, S., & DIRVEN, R. (1997) : *The language of emotions: conceptualization, expression, and theoretical foundation*. Amsterdam, Philadelphia: John Benjamins.
- OATLEY, K. (1989) : The Importance of Being Emotional, *New Scientist*, 123 (Pt. 1678), p. 33-36.
- OHALA, J. J. (1983) : Cross-Language Use of Pitch: An Ethological View, *Phonetica*, 40, p. 1-18.
- OHALA, J. J. (1984) : An ethological perspective on common cross-language utilization of F0 of voice, *Phonetica*, 41, p. 1-16.
- OHALA, J. J. (1994) : The frequency code underlies the sound symbolic use of voice pitch, in *Sound symbolism* (L. Hinton, J. Nichols, & J. J. Ohala, eds.), p. 325-347. Cambridge : Cambridge University Press.
- ORTONY, A., & TURNER, T. J. (1990) : What's Basic about Basic Emotions?, *Psychol. Rev.*, 97, p. 315-331.
- PAKOSZ, M. (1982) : Intonation and Attitude, *Lingua*, 56, p. 153-178.
- PAKOSZ, M. (1983) : Attitudinal Judgements in Intonation: Some Evidence for a Theory, *J. Psycholinguist. Res.*, 12, p. 311-326.
- PITTAM, J., & SCHERER, K. R. (1993) : Vocal expression and communication of emotion, in *Handbook of emotions* (M. Lewis & J. M. Haviland, eds.), p. 185-198. New York: Guilford Press.
- PLOOG, D. (1970) : Social communication among animals, in *The Neurosciences, Second Study Program* (F. O. Schmitt, ed.), p. 349-361. New York: Rockefeller Press.
- ROBINSON, B. W. (1976) : Limbic influences on human speech, *Ann. N.Y. Acad. Sci.*, 280, p. 761-771.
- ROY, D., & PENTLAND, A. (1996) : Automatic Spoken Affect Classification and Analysis, *Proc. Second Int. Conf. Automatic Face Gesture Recognition*, p. 363-367. Los Alamitos, CA: IEEE Comput. Soc. Press.
- RUCH, W. (1987, June) : Personality aspects in the psychobiology of humor laughter. Paper presented at the *third meeting of the International Society for the Study of Individual Differences*, Toronto.

- RUSSELL, J. A. (1991) : Culture and the categorization of emotions, *Psychological Bulletin*, 110, p. 426-450.
- SCHACHTER, S. & SINGER, J. E. (1962) : Cognitive, social and physiological determinants of emotional states, *Psychological Review*, 69, p. 379-399.
- SCHERER, K. R. (1981) : Speech and Emotional States, in *Speech Evaluation in Psychiatry* (J. K. Darby, ed.). New York : Grune and Stratton.
- SCHERER, K. R. (1984a) Emotion as a Multicomponent Process: A Model and some Cross-Cultural Data, *Rev. Personal. Social Psychol.*, 5, p. 37-63.
- SCHERER, K. R. (1984b) : On the nature and function of emotion : a component process approach, in *Approaches to emotion* (K. R. Scherer, P. Ekman, eds.), p. 293-317. Hillsdale, N.J.: Erlbaum.
- SCHERER, K. R. (1985) : Vocal affect signalling: A comparative approach, in *Advances in the study of behavior* (J. Rosenblatt, C. Beer, M. Busnel, & P. J. B. Slater, eds.), p. 189-244. New York: Academic Press.
- SCHERER, K. R. (1988) : On the symbolic functions of vocal affect expression, *Journal of Language and Social Psychology*, 7, p. 79-100.
- SCHERER, K. R. (1994) : Affect bursts, in *Emotions: Essays on emotion theory* (S. H. M. van Goozen, N. E. von de Poll, & J. A. Sergeant, eds.), p. 161-196. Hillsdale, N.J.: Erlbaum.
- SCHERER, K. R. (1995) : Expression of emotion in voice and music, *J. Voice*, 9 (3), p. 235-248.
- SCHERER, K. R., & OSHINSKY, J. (1977) : Cue utilization in emotion attribution from auditory stimuli, *Motiv. Emot.*, 1, p. 331-346.
- SCHERER, K. R., & WALLBOTT, H. G. (1994) : Evidence for universality and cultural variation of differential emotion response patterning, *Journal of Personality and Social Psychology*, 66, p. 310-328.
- SCHERER, K. R., BANSE, R., WALLBOTT, H. G., & GOLDBECK, T. (1991) : Vocal cues in emotion encoding and decoding, *Motivation and Emotion*, 15, p. 123-148.
- SCHERER, K. R., WALLBOTT, H. G., & SUMMERFIELD, A. B., eds. (1986) : *Experiencing emotion: A crosscultural study*. Cambridge, England: Cambridge University Press.
- SPERRY, R. W., GAZZANIGA, M. S., & BOGEN, J. E. (1969) : Interhemispheric relationships : The neocortical commissures ; syndromes of their disconnection, in *Handbook of Clinical Neurology*, vol. 4 (P. J. Vinken & G. W. Bruyn, eds.), p. 273-290. Amsterdam, North Holland.
- SPERRY, R., ZAIDEL, E., & ZAIDEL, D. (1979) : Self recognition and social awareness in the disconnected minor hemisphere, *Neuropsychologia*, 17, p. 153-166.
- TISCHER, B. (1993) : *Die vokale Kommunikation von Gefühlen*. Weinheim: Psychologie Verlags Union / Beltz.
- TOLKMITT, F., BERGMANN, G., GOLDBECK, TH., & SCHERER, K. R. (1988) : Experimental studies on vocal communication, in *Facets of emotion: Recent research* (K. R. Scherer, ed.), pp. 119-138. Hillsdale, N.J.: Lawrence Erlbaum.
- TOMKINS, S. S. (1962) : *Affect, imagery, consciousness, Vol. 1, The positive affects*. New York : Springer.
- VELTEN, E. (1968) : A laboratory task for induction of mood states. *Behaviour Research and Therapy*, 6, p. 473-482.
- WILLIAMS, C. E., & STEVENS, K. N. (1972) : Emotions and Speech: Some Acoustical Correlates, *J. Acoust. Soc. Am.*, 52, p. 1238-1250.
- WUNDT, W. (1913) : *Grundriss der Psychologie*. Leipzig : Alfred Kröner.

2. Articles qui ne sont pas directement liés aux émotions

(modèles de l'intonation basés sur des contours ; synthèse de la parole ; autres)

a) Articles que nous avons lu

- AUBERGÉ, V. (1992) : Developing a structured lexicon for synthesis of prosody, in *Talking Machines: Theories, Models, and Designs* (G. Bailly, C. Benoît, & T. R. Sawallis, eds.), p. 307-321. Amsterdam: Elsevier Science Publishers.

- AUBERGÉ, V., & BAILLY, G. (1995) : Generation of intonation: a global approach, *Eurospeech 95, Madrid*, p. 2065-2068.
- GELUCK, P. (1997) : *Le chat. Pensées blanches*. Casterman.
- MÖBIUS, B. (1995) : Components of a quantitative model of German intonation, *Proceedings of the 13th ICPhS, Stockholm*, Vol. 2, p. 108-115.
- MORLEC, Y., AUBERGÉ, V., & BAILLY, G. (1995) : Evaluation of Automatic Generation of Prosody with a Superposition Model, *Proceedings of the 13th ICPhS, Stockholm*, Vol. 4, p. 224-227.
- MORLEC, Y., BAILLY, G., & AUBERGÉ, V. (1996) : Un modèle connexioniste modulaire pour l'apprentissage des "gestes" intonatifs, *21es journées d'étude sur la parole*, p. 207-210. Avignon.
- MORLEC, Y., BAILLY, G., & AUBERGÉ, V. (1997) : Apprentissage Automatique d'un Module de Génération Multistyle de l'Intonation, 1^{er} JST Francil, p. 407-412. Avignon.
- MURPHY, K., CORFIELD, D. R., FINK, G. R., WISE, R. J. S., GUZ, A., & ADAMS, L. (1997) : Neural mechanisms associated with the control of speech in man, *NeuroImage*, Vol. 5 (4), S253.
- ROCK, I., & PALMER, S. (1991) : L'héritage du gestaltisme, *Pour la science*, 160, Fév. 91, p. 64-70.
- VAISSIÈRE, J. (1983) : Language-Independent Prosodic Features, in *Prosody: Models and Measurements* (A. Cutler & D. R. Ladd, eds.), p. 53-66. Berlin, Heidelberg, New York, Tokyo : Springer-Verlag.

b) Articles intéressants, que l'on pourrait lire

- GIBBON, D., & RICHTER, H. (1984) : *Intonation, Accent and Rhythm: studies in discourse phonology*. Berlin, New York: W. de Gruyter.
- GREENE, B. G., LOGAN, J. S., & PISONI, D. B. (1986) : Perception of Synthetic Speech Produced Automatically by Rule: Intelligibility of Eight Text-to-Speech Systems, *Behav. Res. Methods Instrum. Comput.*, 18, p. 100-107.
- LARKEY, L. S. (1983) : Reiterant speech: An acoustic and perceptual validation, *J. Acoust. Soc. Am.*, 73 (4), p. 1337-1345.
- LIBERMAN, M., & PIERREHUMBERT, J. (1984) : Intonational Invariance Under Changes in Pitch Range and Length, in *Language Sound Structure* (M. Aronoff, R. Oehrle, eds.). Cambridge, Massachusetts : MIT Press.
- LIBERMAN, M. Y., & STREETER, L. A. (1978) : Use of nonsense-syllable mimicry in the study of prosodic phenomena, *Journal of the Acoustic Society of America*, 63 (1), p. 231-233.
- MÖBIUS, B. (1997) : Synthesizing German Intonation Contours, in *Progress in Speech Synthesis* (J. Van Santen, R. Sproat, J. Olive, & J. Hirschberg, eds.), p. 401-416. New York: Springer.
- MORLEC, Y., BAILLY, G., & AUBERGÉ, V. (1996) : Generating Intonation by Superposing Gestures, *Proc. ICSLP 96, Philadelphia*.
- MORLEC, Y., BAILLY, G., & AUBERGÉ, V. (1997) : Synthesizing attitudes with global rhythmic and intonation contours, *Eurospeech 97, Athen*.
- ROSSI, M., DI CHRISTO, A., HIRST, D., MARTIN, PH., NISHINUMA, Y. (1981) : *L'intonation, de l'acoustique à la sémantique*. Paris : Klincksieck.

INDEX

A		N	
aire motrice supplémentaire	<i>Voir supplementary motor area</i>	numérisation	48
anosognosie	11		
C		O	
code fréquentiel	<i>Voir frequency code</i>	ontogenèse	iv
component process model	5; 20	orbicularis oculi	28
cortex préfrontal	7	ordre de présentation	51
E		P	
EEG	29	perception de l'état d'arrière-plan du corps	10
émotions primaires	6	phylogenèse	iv; 3
émotions secondaires	7	présélection	36
établissement de corpus	33; 43		
éthologie	iv; 3; 15	Q	
expression attitudinale	16	questionnaires	37
F		R	
Facial Action Coding System	29	réalisations type	40
FACS	<i>Voir Facial Action Coding System</i>	registre	20
fonctions des émotions	3	resynthèse	19
frequency code	16	rire	29
H		S	
hémisphère droit	8	SEC	<i>Voir stimulus evaluation check</i>
histoires cadres	34	signal	15; 16; 30
L		simulation de la perception des émotions	12
laugh-speak	30	SMA	<i>Voir supplementary motor area</i>
locuteurs	46	sourire	8; 27
M		sourire de Duchenne	9; 28
matériel d'enregistrement	44	sourire mécanique	41; 47
matrice de confusion	iv; 22; 25; 60	stimulus evaluation check	5
meta-connaissances	37	supplementary motor area	31
mise en situation	43	symbole	15; 16
modèle de configurations	17	symptôme	15
modèle de covariance	16	système limbique	6
modèle des canaux parallèles	16	T	
modèle des processus composants	<i>Voir component process model</i>	test de perception	37; 51
model		test de signification	iv; 39
V		V	
		vocalisations de singes	24