UNIVERSITE DE GENEVE
SECTION DE PSYCHOLOGIE

FACULTE DE PSYCHOLOGIE ET
DES SCIENCES DE L'EDUCATION

COMMUNICATION VOCALE DES ÉMOTIONS

Perception de l'expression vocale et attributions émotionnelles

- THÈSE -

Sous la direction du Professeur
Klaus R. Scherer

Présentée à la
Faculté de Psychologie et des Sciences de l'Education
de l'Université de Genève
en vue de l'obtention du grade de Docteur en Psychologie
par

Tanja Bänziger

Thèse N° 325

Genève 2004

Remerciements

Cette thèse a été réalisée grâce au soutien d'un grand nombre de personnes qui m'ont conseillée et encouragée durant plusieurs années.

En premier lieu, je tiens à remercier mon directeur de thèse – Klaus Scherer – dont les travaux ont très largement inspiré les études effectuées dans cette thèse. Mes remerciements s'adressent également aux membres de ma commission et de mon jury de thèse: Johan Sundberg, Eric Wehrli, Ulrich Frauenfelder et Véronique Aubergé. Je les remercie pour la (re)lecture du manuscrit, pour leurs critiques et leurs suggestions.

Cette thèse a également bénéficié de l'aide et des conseils de plusieurs experts dans des domaines variés de l'analyse de la prosodie. Je tiens à remercier tout particulièrement Bob Ladd, pour ses propositions relativement à la stylisation de la F0; Piet Mertens, pour ses conseils et son soutien dans le cadre du Projet Plurifacultaire Prosodie; Antoine Auchlin, également pour ses conseils et tout particulièrement pour sa participation vocale; et finalement Michel Morel pour la réalisation des expressions de synthèse.

Mes remerciements vont également à tous mes collègues – doctorants ou assistants – qui se sont succédés au sein du groupe de recherche sur les émotions; avec une mention spéciale pour Tom Johnstone qui a contribué à l'orientation de mon travail dans sa phase initiale. Stéphanie Meylan, Véronique Tran et Ofra Hazanov ont contribué à la réalisation de ce travail aussi bien par leur soutien professionnel que par leur soutien affectif. David Sander et Didier Grandjean ont supporté et (supportent encore) des intrusions répétées dans leur bureau pour des questions de toutes sortes. J'espère que tous les autres collègues passés ou présents, (Carien, Janique, Gudrun, Tatjana, Virginie, …) me pardonneront de ne pas les avoir également cité(e)s, ces remerciements leurs sont également destinés.

Je suis également particulièrement reconnaissante à Ursula Scherer qui est à l'origine de mon engagement dans l'équipe de recherche sur les émotions. Sans son intervention, ce travail n'aurait jamais vu le jour.

Finalement, mes remerciements vont surtout à mes parents – qui m'ont toujours soutenue moralement dans mes études – et à Anders, son soutien et ses encouragements ont été inépuisables.

Tables des matières

1 Introduction

> *It has often struck me as a curious fact that so many shades of expression are instantly recognized without any conscious process of analysis on our part. No one, I believe, can clearly describe a sullen or sly expression; yet many observers are unanimous that these expressions can be recognized in the various races of man. […] So it is with many other expressions, of which I have had practical experience in the trouble requisite in instructing others what points to observe.*
>
> Darwin, 1872
> édité par Ekman, 1998, p. 355

Charles Darwin est le premier auteur qui s'est intéressé à l'étude systématique d'expressions correspondant à des émotions telles que la colère, la joie, la tristesse ou la peur. Ses observations se sont portées aussi bien sur les expressions vocales que sur les expressions faciales, produites par des êtres humains appartenant à différentes cultures, ainsi que par des animaux de différentes espèces. Le 20ème siècle a vu se développer considérablement l'étude des expressions faciales pour lesquelles des systèmes de codage qui permettent de décrire avec précision tous les mouvements du visage sont aujourd'hui disponibles (par exemple le Facial Action Coding System d'Ekman et Friesen, 1978). Comparativement, l'étude des expressions vocales a été plus longtemps négligée. Le regain d'intérêt rencontré depuis quelques années par ce domaine d'étude laisse cependant présager d'un développement, comparable à celui réalisé dans le domaine des expressions faciales, des méthodes utilisées pour l'étude des expressions vocales. Les méthodes qui ont été et qui sont encore actuellement utilisés dans ce domaine d'étude seront présentés en détails ci-dessous dans la seconde partie de cette introduction.

La citation de Darwin, reproduite ci-dessus, s'applique encore aujourd'hui à deux aspects de l'étude de la communication vocale des émotions. Premièrement, elle reflète la problématique centrale de ce domaine d'étude : les émotions exprimées par la voix peuvent être reconnues sans difficulté par la plupart des individus, sans que ces individus soient nécessairement capables d'indiquer sur quels aspects des expressions ils basent leurs attributions. Deuxièmement, elle reflète aussi l'état actuel de la recherche sur les expressions vocales émotionnelles : dans ce domaine, de nombreuses études ont démontré que les expressions vocales permettent d'identifier l'état émotionnel d'un individu en l'absence d'autres indices; en revanche, les caractéristiques vocales qui permettent aux auditeurs d'identifier correctement l'émotion exprimée restent encore mal connues. Les principaux résultats obtenus dans les études effectuées relativement à la reconnaissance des expressions vocales et aux

caractéristiques vocales qui correspondent à différentes émotions exprimées seront décrits ci-dessous dans la seconde partie de l'introduction.

Mais avant d'entreprendre cet état des lieux relatif aux méthodes utilisées dans ce domaine ainsi qu'aux résultats déjà obtenus, quelques aspects théoriques méritent également d'être introduits. La conceptualisation des réactions émotionnelles et des expressions qui leurs sont associées n'est évidemment pas étrangère aux opérationalisations utilisées pour étudier la communication vocale des émotions; et certaines notions théoriques – telles que l'intensité et l'activation émotionnelle – interviennent directement au niveau des interprétations réalisées par les auteurs qui ont effectué des travaux dans ce domaine.

1.1 Aspects théoriques

Plusieurs courants théoriques proposent aujourd'hui différents modèles et définitions des émotions (Cornelius, 1996). Une définition absolue de l'Emotion ne peut donc être formulée. Cette situation – partagée par beaucoup d'autres concepts psychologiques – ne doit cependant pas nous empêcher de présenter et de discuter un certain nombre de points théoriques, qu'ils soient consensuels ou controversés. L'objectif de cette section relative aux *aspects théoriques* n'est toutefois pas de présenter et comparer différentes perspectives théoriques; elle vise, plus simplement à introduire un certain nombre de points théoriques directement liés à la problématique de l'expression et de la communication vocale des émotions. Dans cette section, les expressions vocales sont, dans un premier temps (point 1.1.1), considérées indépendamment de toute régulation et, par la suite (point 1.1.2), dans la perspective de la régulation introduite par la communication des états émotionnels.

1.1.1 L'expression vocale comme composante de la réaction émotionnelle

De manière consensuelle, l'émotion est définie comme une *réaction* (à un événement/situation réel/le ou imaginaire) comprenant *plusieurs facettes ou composantes*. Il existe un désaccord assez important relativement au nombre de ces composantes et à leur importance respective dans la définition de la réaction émotionnelle. Toutefois, *trois composantes* sont généralement acceptées comme constituants essentiels de la réaction émotionnelle. Il s'agit du *sentiment subjectif* (vécu émotionnel, "feeling" en anglais), de la *réaction physiologique* (plus ou moins différenciée selon les modèles théoriques) et de l'*expression émotionnelle* (faciale, vocale ou posturale). Dans les paragraphes qui suivent, différents aspects de la relation théorique entre l'expression émotionnelle (en particulier l'expression vocale) et les deux autres composantes – soit le sentiment émotionnel et la réaction physiologique – seront examinés, dans une perspective qui vise à mettre en évidence quelques problématiques et concepts centraux dans le domaine de l'expression vocale des émotions.

1.1.1.1 Expressions vocales et réactions physiologiques – activation émotionnelle

La relation entre l'expression vocale et la réaction physiologique associées à une réaction émotionnelle est particulièrement importante sur le plan théorique. Il est en effet aisé de concevoir qu'un certain nombre de modifications physiologiques périphériques associées à des réactions émotionnelles puissent affecter directement le système de production vocale et en conséquence induire des modifications sur le plan de l'expression vocale. Dans une revue publiée en 1986, Scherer a formulé un ensemble de prédictions concernant les effets sur la voix attendus pour différents changements de l'état émotionnel d'un locuteur. Dans ce modèle, les changements physiologiques associés à l'état émotionnel d'un locuteur affectent l'ensemble du système de production vocale; le fonctionnement des appareils respiratoire, phonatoire et articulatoire est modifié par les changements autonomes et somatiques associés à la réaction émotionnelle (v. également Scherer, 2003).

Cette conception suppose l'existence de réactions physiologiques périphériques (autonomes et/ou somatiques) spécifiques à différents états émotionnels, en l'absence desquelles les expressions vocales (spontanées, non-régulées) correspondant à différents états émotionnels ne seraient pas différenciées. Depuis la proposition originale de James (1884), la spécificité des réactions physiologiques périphériques associées à différentes émotions a été défendue par différents auteurs. Dans la perspective de Scherer (1986, 2003) évoquée ci-dessus, différentes dimensions de l'évaluation cognitive à l'origine de la réaction émotionnelle – telles que la pertinence, l'agrément ou le degré de contrôle évalués dans la situation source d'émotion – synchronisent les différentes composantes (physiologique, expressive et subjective) pour donner lieu à un état émotionnel différencié comprenant une réaction physiologique, une réaction expressive et une réaction subjective ("feeling") spécifiques. Les modèles qui défendent l'existence d'un nombre limité d'émotions fondamentales ("basic emotions", Ekman, 1992) ont par ailleurs postulé et tenté de démontrer l'existence de patterns physiologiques périphériques spécifiques pour un certain nombre "d'émotions fondamentales" (v. par exemple Ekman, Levenson, & Friesen, 1983).

La question de la spécificité des réactions physiologiques périphériques associées à différents états émotionnels a cependant été mise en cause par d'autres auteurs qui ont affirmé que sur le plan périphérique, seul un niveau d'*activation* physiologique indifférencié est associé à toute réaction émotionnelle (Schachter & Singer, 1962; Gray, 1987). Par extension, ce niveau d'activation peut également être conçu comme une dimension sous-jacente aux réactions émotionnelles (v. Russell, 1980, pour une illustration de l'utilisation de l'activation comme dimension sous-jacente aux catégories émotionnelles). Dans cette perspective, une activation physiologique ("arousal") très

importante pourrait être associée à certaines émotions (par exemple la peur panique) alors que d'autres émotions (par exemple la tristesse déprimée) comporteraient un niveau d'activation beaucoup plus faible. Bien évidemment, dans cette optique, l'influence de la physiologie périphérique sur les expressions vocales se limite également à refléter un niveau d'activation global. Cette dernière hypothèse est régulièrement évoquée dans le cadre de l'interprétation des résultats obtenus par les études qui se sont intéressées à décrire les caractéristiques acoustiques des expressions vocales émotionnelles. En conséquence, la question de l'influence de l'activation physiologique sur les expressions vocales sera développée plus longuement ci-dessous (sections 1.1.1.2 et 1.2.1.2).

1.1.1.2 Expressions vocales et sentiment subjectif – valence et intensité émotionnelle

Le sentiment subjectif ("feeling") associé à la réaction émotionnelle intervient également de manière plus ou moins directe dans le cadre de l'étude des expressions vocales. Certains concepts initialement associés au sentiment émotionnel telle que la *valence* et l'*intensité* sont notamment régulièrement évoqués dans ce domaine. Ces concepts trouvent leur origine dans la démarche introspective qui vise à décrire les sentiments émotionnels dans une perspective phénoménologique. La première proposition dans ce sens a été formulée par Wundt (1897) qui a suggéré que les états émotionnels pouvaient être décrits, dans une perspective subjective, en fonction de leur positions respectives sur trois dimensions: la valence (état perçu comme positif/plaisant - négatif/déplaisant), l'excitation (état perçu comme calme - excité) et la tension (état perçu comme relaxé - tendu). Par la suite, des dimensions alternatives ont été proposées par d'autres auteurs (par exemple Schlossberg, 1954; Osgood, 1966 ou Plutchik, 1962). Malheureusement, il existe encore actuellement une certaine confusion au sujet de l'utilisation et de la définition d'un certain nombre de ces dimensions. Deux aspects, en particulier, sont souvent utilisés et, à notre avis, insuffisamment explicités: la notion de réaction (et d'expression) *neutre*, d'une part, et la relation entre l'activation (évoquée dans la section précédente) et l'*intensité* émotionnelle, d'autre part.

Le concept d'émotion *neutre* est relativement problématique. La *neutralité* fait en principe référence à la dimension bipolaire de *valence* perçue, une émotion *neutre* serait une émotion située au centre de cette dimension, c'est-à-dire une émotion vécue comme ni négative/déplaisante, ni positive/plaisante. Or la plupart des définitions de l'émotion insistent sur la coloration affective positive/négative des réactions émotionnelles. Dans cette optique (si l'on adopte une telle définition) un état *neutre* ne pourrait être considéré comme un état émotionnel. Indépendamment de cette question, un état *neutre* qualifie donc un état non-positif et non-négatif; en revanche ce qualificatif ne définit, en principe, pas les autres dimensions qui pourraient être également utilisées pour

qualifier cet état. Plus concrètement, dans le domaine de l'étude des expressions vocales, les expressions *neutres* correspondent par définition à un état de *valence neutre*, en revanche l'*activation*, la *tension* ou encore le degré de *contrôle* associé à cet état ne sont pas explicitement définis; bien qu'en pratique, il s'agit généralement d'un état qui comprend un niveau d'*activation* et de *tension faible* et un niveau de *contrôle élevé*. A moins de faire l'hypothèse que seule la valence de l'état émotionnel importe dans le domaine des expressions vocales émotionnelles, il n'est donc pas anodin d'utiliser des expressions "neutres" comme point de référence et de leur comparer indistinctement tous les autres types d'expressions émotionnelles. A notre avis, il est nécessaire dans ce cas d'expliciter également les caractéristiques particulières (par exemple un niveau de contrôle élevé ou une activation faible) des expressions "neutres" qui ne sont réellement "neutres" que sur le plan de la valence du sentiment subjectif.

Une autre source de confusion réside dans l'équivalence conceptuelle, à notre avis infondée, qui est souvent réalisée pour les concepts d'*activation* et d'*intensité*. Le terme intensité désigne habituellement l'importance de l'émotion sur le plan du vécu subjectif. Une émotion intense est une émotion vécue comme forte, une émotion peu intense est une émotion ressentie comme faible (pour plus de détails sur le concept d'intensité émotionnelle: v. Guerrero, Andersen, & Trost, 1998; Frijda, Ortony, Sonnemans, & Clore, 1992). Le concept d'activation, tel qu'il a été introduit ci-dessus, désigne en premier lieu l'activation physiologique (autonome et somatique) associée à l'émotion. Cette dimension se reflète également sur le plan du vécu subjectif sous forme d'un niveau d'activation/excitation perçue dans le cadre d'une réaction émotionnelle. Certains modèles/définitions de l'émotion postulent une relation forte entre ces deux dimensions (*intensité* et *activation*). La tradition théorique qui place les réactions périphériques (physiologiques, mais également motrices et donc expressives) au centre de la réaction émotionnelle trouve ses origines dans les propositions de James (1884). Le courant le plus conservateur de cette tradition soutient que le sentiment émotionnel résulte de la perception (proprioception) d'un pattern d'activation périphérique spécifique (une configuration de réactions autonomes et motrices). Une conception plus modérée, et plus largement partagée, issue de cette tradition théorique, postule une relation entre la présence d'une réaction périphérique (autonome et/ou motrice) et l'*intensité* du sentiment émotionnel. Dans cette perspective, une activation périphérique faible serait associée à une émotion vécue comme faible, alors qu'une activation périphérique forte serait associée à une émotion vécue comme forte. A notre avis ce postulat théorique ne doit toutefois pas résulter dans une confusion complète de ces deux dimensions. En particulier dans le domaine des expressions vocales où le niveau d'activation sous-jacent à la réaction émotionnelle joue probablement un rôle très important,

l'existence potentielle d'états émotionnels vécus comme très intenses mais associés à un niveau d'activation physiologique relativement faible ne doit pas être exclue a priori (Guerrero et al., 1998).

1.1.1.3 Perception des expressions vocales et sentiment subjectif – imitation et feedback

Un prolongement de la perspective théorique évoquée ci-dessus (le modèle "périphéraliste") a donné lieu à une hypothèse relative à la manière dont les expressions émotionnelles (faciales ou vocales) sont perçues. Cette hypothèse a été formulée par Lipps (1903) et développée notamment par Hatfield, Cacioppo, & Rapson (1994). Ces auteurs ont proposé que la reconnaissance des émotions exprimées par autrui est basée sur les expériences émotionnelles vécues par les observateurs. Dans cette perspective, un observateur "reproduit" (plus ou moins consciemment et ouvertement) les expressions qu'il observe et se base sur la sensation interne associée à cette expression (le sentiment subjectif) pour "reconnaître" l'émotion exprimée par autrui (Hess & Blairy, 2001 ; Wallbott, 1991). Ce mécanisme s'apparente d'avantage à un phénomène d'empathie qu'à une reconnaissance abstraite, l'émotion d'autrui est *ressentie* (suite à "l'imitation" des expressions observées) plus qu'elle n'est *reconnue*.

Cette reconnaissance des expressions par imitation et feedback proprioceptif est parfois opposée à un modèle plus classique d'apprentissage par associations. Dans cette perspective, les émotions exprimées sont reconnues par un observateur qui analyse (plus ou moins consciemment) les indices vocaux/acoustiques ou faciaux/visuels disponibles et décide, sur la base d'observations antérieures, à quel type d'émotion ces indices correspondent. Dans cette perspective, des associations entre des combinaisons d'indices vocaux/ou faciaux et des réactions émotionnelles globales, ainsi que les caractéristiques des situations qui les déclenchent sont formées au cours de l'histoire individuelle de l'observateur et suffisent à rendre compte de la reconnaissance, sans faire appel aux réactions émotionnelles vécues par l'observateur.

Il importe de relever que ces deux mécanismes ne sont en réalité pas exclusifs. Dans la plupart des cas, les deux processus (reconnaissance abstraite et imitation/feedback) contribuent probablement conjointement à la reconnaissance des émotions exprimées.

1.1.2 L'expression vocale régulée dans (par) la communication

Les expressions (faciales et vocales) sont la manifestation observable des réactions émotionnelles. Ce statut leur confère une double fonctionnalité dans la communication et la gestion des interactions sociales. Les expressions émotionnelles permettent, en principe, aux émetteurs (de ces expressions) d'influencer le comportement de leurs interlocuteurs et permettent aux interlocuteurs de formuler des prédictions relativement aux comportements à venir des émetteurs.

Dans ce contexte de communication, les expressions sont bien évidemment régulées (modulées ou transformées) en fonction d'un ensemble de règles socioculturelles. Ces règles ("display rules") ont été étudiées par différents auteurs qui ont montré que les expressions émotionnelles varient d'une culture (ou d'un groupe social) à l'autre, dans des contextes "objectivement" similaires, en fonction de règles intégrées par les membres de ces cultures (groupes sociaux) relativement à ce qu'il convient de montrer ou de ne pas montrer. (v. notamment Ekman & Friesen, 1969; Matsumoto, 1990; LaFrance & Hecht, 1999). Les régulations induites par le contexte social peuvent être de différents types. Certaines situations sociales peuvent exiger la *suppression* de certaines expressions alors que d'autres situations au contraire exigent de *montrer* ou même d'*exagérer* des expressions spécifiques. Des expressions peuvent être utilisées également de manière à *masquer* une expression spontanée qui ne serait pas désirable dans un contexte social donné.

En parallèle aux règles sociales qui sont appliquées, souvent inconsciemment et automatiquement, dans la plupart des interactions sociales. Le contexte social peut également influencer les expressions émotionnelles en fonction de la position et des objectifs de l'émetteur dans la situation. Dans une situation de communication un individu peut utiliser ses expressions émotionnelles de manière à influencer – plus ou moins in/consciemment et plus ou moins in/volontairement – les réactions des ses interlocuteurs. Un exemple de ce type de régulation a été récemment donné par Reissland, Shepherd, & Cowie (2002) qui ont observé un ajustement des expressions vocales d'un groupe de jeunes mères relativement aux expressions faciales de leurs bébés. Dans cette étude, les mères et leurs bébés jouaient à un jeu d'élicitation de surprise (avec un diable à ressort); les mères ont produit des expressions vocales dont la fréquence fondamentale était d'autant plus élevée que l'expression de surprise (désirée) chez leur enfant était faible. Ce résultat peut être interprété comme correspondant à une exagération de l'expression de surprise chez la mère destinée à stimuler (renforcer) l'expression de surprise chez le bébé.

En outre, les expressions vocales sont également affectées par un ensemble de facteurs non-émotionnels. Les influences linguistiques et paralinguistiques sur les expressions vocales ne sont notamment pas négligeables. Ces influences incluent en particulier les phénomènes d'emphase dans le discours (destinés à accentuer certains éléments) et les actes pragmatiques, tels que les expressions de doute ou de désapprobation ou encore les interrogations qui peuvent être communiquées par des modifications de l'intonation ou de la qualité vocale. De plus, différents aspects propres au locuteur influencent également les expressions vocales. Certaines inflexions vocales peuvent par exemple être attribuées à un accent régional, à la personnalité, au style expressif, à l'âge ou encore à l'état de santé du locuteur.

Dans le cadre de cette thèse, notre intérêt se limitera exclusivement aux expressions émotionnelles. Il importe toutefois de souligner que ces expressions sont en pratique difficiles à isoler des autres composantes expressives. De plus, comme indiqué ci-dessus, les expressions émotionnelles sont elles-mêmes pratiquement toujours soumises à des régulations qui dérivent de règles sociales ou de stratégies de communication. Des expressions émotionnelles "pures" ou "spontanées" surviennent donc probablement assez rarement dans la vie quotidienne; et, dans la mesure où il est évidemment très difficile pour les chercheurs d'enregistrer de telles expressions, elles sont également très rarement étudiées. Afin de contrôler les différences expressives interindividuelles, la méthode la plus fréquemment utilisée pour étudier les expressions vocales émotionnelles consiste en effet à enregistrer des acteurs qui simulent un nombre prédéfini de réactions émotionnelles (pour plus de détails v. section 1.2 ci-dessous). Cette méthode a été souvent critiquée du fait que les expressions ne correspondraient pas ou peu à des expressions émotionnelles "véritables" ("pures" ou "spontanées"), mais plutôt à des modèles enseignés dans les cours d'art dramatique. En réalité, il est peu probable que ces expressions ne correspondent en rien aux expressions émotionnelles qui surviennent en situation sociale dans la vie quotidienne. Afin de paraître crédibles et d'avoir un impact sur leurs auditeurs, l'intérêt des acteurs est d'utiliser des codes d'expressions qui seront interprétés comme authentiques. En revanche, l'exagération des codes sociaux de communication est certainement présente dans les enregistrements réalisés par des acteurs (v. également Banse & Scherer, 1996, pour une discussion à ce propos).

1.1.3 Catégories et/ou dimensions émotionnelles pour l'étude des expressions vocales

Le dernier aspect théorique abordé ci-dessous concerne l'utilisation de catégories émotionnelles ou de dimensions émotionnelles dans le cadre de l'étude des expressions vocales. Dans ce domaine, la plupart des études effectuées ont utilisé des catégories émotionnelles – telles que la 'peur', la 'joie', le 'dégoût', etc. – pour identifier les états émotionnels théoriquement sous-jacents aux expressions étudiées. Les mêmes catégories sont également utilisées dans les études destinées à évaluer la reconnaissance des émotions exprimées (v. section 1.2 ci-dessous). Au-delà des labels qui les identifient, ces catégories ne sont en général pas définies théoriquement. Ce manque de spécification sur le plan des définitions pose évidemment un certain nombre de problèmes. Les catégories identifiées par les labels émotionnels les plus fréquemment utilisés dans ce domaine (colère, peur, joie et tristesse) sont relativement larges et sont donc susceptibles de correspondre à un éventail plus ou moins large d'états émotionnels différents; la colère, par exemple, pourrait correspondre à un état légèrement irrité ou à une rage intense; la peur à de l'inquiétude ou de la peur panique. Très peu d'études se sont intéressées à la possibilité de différencier des expressions vocales

correspondant à différents états émotionnels appartenant à une même "famille" (identifiée par un label tel que colère, tristesse, peur ou joie). Cet état de fait est probablement dû à l'influence dominante du courant théorique qui défend l'existence d'un nombre limité "d'émotions de base" (ou "émotions fondamentales"). Selon ce courant théorique, un pattern expressif défini correspond à chaque "émotion de base". Le nombre et la qualité des "émotions de base" varient en fonction des différents auteurs qui représentent ce courant théorique; bien qu'il semble exister un consensus sur au moins un sous-ensemble "d'émotions fondamentales" qui correspondraient à la joie, la peur, la tristesse, la colère, le dégoût et la surprise.

Deux études, au moins, ont cependant montré que des expressions correspondant à une même "émotion de base" peuvent être effectivement différentes: (1) Frick (1986) a distingué un type de 'colère' lié à l'agression et un type de 'colère' lié à la frustration. Ses résultats indiquent que ces deux types de 'colère' se différencient sur le plan acoustique et démontrent également que des juges peuvent discriminer correctement un ensemble d'expressions vocales 'agressives' relativement à des expressions 'frustrées' ; (2) Banse & Scherer (1996) ont mis en évidence des profils acoustiques différents à l'intérieur des "familles" émotionnelles 'peur', 'tristesse', 'joie' et 'colère'. Dans une large mesure (avec quelques confusions), les états correspondant à une même "famille" ont également été différenciés par un groupe d'auditeurs dans cette étude.

Ces résultats confirment donc que différents états d'une même "famille" émotionnelle (une catégorie large identifiée par un label d'ordre général) peuvent être différenciés sur le plan des expressions vocales qui leur correspondent. Le fait de ne pas spécifier l'état désigné par un label d'ordre général a notamment pour conséquence de limiter la possibilité de comparer les résultats obtenus dans différentes études qui utilisent les mêmes labels (sans en préciser la définition). Une plus grande précision sur le plan de la définition des états émotionnels, et donc des expressions considérées, est en conséquence nécessaire. Une méthode parfois utilisée pour préciser la définition des états émotionnels consiste à utiliser des scénarios, des descriptions plus ou moins détaillée de situations à l'origine des états émotionnels et des expressions étudiées (v. par exemple Banse & Scherer, 1996).

Une autre possibilité consiste à définir les états émotionnels sous-jacents aux expressions utilisées sur un ensemble de dimensions continues qui permettent de différencier un grand nombre d'états émotionnels différents. Cette perspective a été notamment développée par (Scherer, 1986, 2003) qui a proposé d'utiliser un modèle de l'évaluation cognitive antécédente aux réactions émotionnelles ("appraisal model") comprenant plusieurs dimensions (ou critères d'évaluation). Dans ce modèle, un grand nombre d'états différents, et d'expressions différentes, peuvent théoriquement résulter de la valeur attribuée par un individu particulier à une situation/événement spécifique sur un ensemble de

dimensions/critères d'évaluation comprenant: (1) la pertinence de l'événement pour la personne, (2) les implications de cet événement relativement aux buts et aux besoins de la personne, (3) le potentiel de maîtrise/contrôle évalué par cet individu relativement à la situation et, finalement, (4) la compatibilité de l'événement avec les normes et standards personnels et sociaux généralement admis par cette personne.

D'autres dimensions peuvent évidemment être invoquées, notamment les dimensions proposées comme descriptives des sentiments subjectifs associés aux émotions, tels que la valence ou l'activation vécues (cf. section 1.1.1.2). L'utilisation de dimensions continues permet de différencier un nombre d'états émotionnels plus important que les états habituellement reconnus comme des "émotions de base" (v. Russell & Feldman Barrett, 1999, pour une argumentation plus poussée concernant "l'avantage" des dimensions sur les catégories fondamentales). Toutefois, les "dimensions essentielles" ne sont pas plus objectivement identifiables que les "catégories de base". La question des dimensions ou des catégories qui seraient particulièrement pertinentes pour l'étude des expressions vocales est donc loin d'être résolue. En conséquence, dans la perspective d'un progrès significatif dans ce domaine de recherche, cette question requiert qu'un nombre plus important de propositions théoriques et de tests empiriques soient rapidement formulées/réalisés dans cette direction (v. Cowie, 2000, pour une discussion à ce sujet).

Dans ce cadre, il est également important de ne pas perdre de vue que les études sur les expressions vocales font le plus souvent appel aux catégories et/ou aux dimensions émotionnelles dont disposent les locuteurs qui produisent les expressions et les auditeurs qui reconnaissent (ou discriminent) les émotions exprimées. Dans cette perspective, un nombre très grand de catégories, dont certaines ne présenteraient peut-être que des différences relativement subtiles, risquerait de ne pas être appréhensible pour des individus sans connaissances spécifiques (sans entraînement ou formation préalable) dans le domaine des réactions et des expressions émotionnelles. Il s'agit donc de formuler une catégorisation ou un espace dimensionnel suffisamment "sophistiqué" pour inclure les états émotionnels et les expressions vocales pertinentes, mais également suffisamment "simple" pour correspondre aux représentations d'individus "naïfs" relativement à cette problématique.

Comme nous l'avons mentionné ci-dessus, il existe dans la littérature quelques indications isolées qui suggèrent que des auditeurs naïfs sont en mesure de différencier des expressions vocales qui correspondent à des états émotionnels plus subtils que des émotions fondamentales – différents types de colère (Frick, 1986) ou différents types de joie, de peur ou de tristesse (Banse & Scherer, 1996). Pourtant, d'autres données suggèrent que les catégories correspondant aux "émotions de base" reflètent assez bien l'organisation des états émotionnels sur le plan des représentations des

individus interrogés à ce sujet. Shaver, Schwartz, Kirson, & O'Connor (1987) ont par exemple défendu l'existence "d'émotions prototypiques" (centrales sur le plan des représentations des individus) qui correspondraient plus ou moins aux "émotions de base" proposées par d'autres théoriciens. Dans le même sens, certains auteurs ont proposé qu'un système dimensionnel relativement rudimentaire pourrait suffire à rendre compte des représentations associées aux émotions. Dans le domaine des expressions vocales, Green & Cliff (1975) ont notamment montré que les dimensions valence et activation permettent de rendre compte d'une large partie des jugements de similarité effectués par un groupe d'auditeurs relativement à des expressions vocales émotionnelles. Ce résultat rejoint les analyses de Russell (1980) qui défend l'existence d'une organisation des représentations émotionnelles dans un espace bidimensionnel définis par la valence et l'activation.

Les différents points exposés ci-dessus révèlent que les problématiques associées à l'étude des expressions vocales émotionnelles sont nombreuses. Malheureusement, elles sont généralement peu formalisées dans le cadre des travaux de recherche qui ont été consacrés aux expressions vocales. Beaucoup d'études dans ce domaine sont en conséquence effectuées dans une perspective essentiellement exploratoire et se fondent assez rarement sur des hypothèses ou des motivations théoriques. La section ci-dessous sera consacrée à une revue des méthodes et des résultats obtenus par ses études.

1.2 Revue des méthodes utilisées et des principaux résultats obtenus

Deux revues de la recherche sur les expressions vocales ont été publiées durant ces derniers mois: Scherer (2003) a, d'une part, réalisé une revue très complète des paradigmes utilisés dans ce domaine; Juslin & Laukka (2003) ont, d'autre part, publié une revue des résultats obtenus par 104 études qui se sont intéressées aux expressions vocales émotionnelles. Bien que ces deux articles offrent déjà un reflet exhaustif de ce domaine d'étude, les principaux paradigmes utilisés (la méthodologie) et les principaux résultats obtenus dans les études de l'expression vocale émotionnelle seront présentés synthétiquement ci-dessous. L'objectif de cette présentation est essentiellement de permettre de situer, ultérieurement, la méthodologie et les questions examinées dans cette thèse relativement aux études effectuées par d'autres auteurs dans ce domaine.

Les recherches effectuées dans le domaine de l'expression et de la communication vocale des émotions peuvent être classées en deux catégories en fonction de leur centre d'intérêt principal:

(1) Certaines études se centrent sur la **production** des expressions vocales, sur les processus d'**encodage** de l'émotion dans la voix. Ces études s'efforcent de décrire l'effet de différents

états émotionnels sur un ensemble de caractéristiques vocales. Des explications concernant la manière dont les émotions produisent ces effets sont assez rarement avancées.

(2) D'autres études se centrent sur la **perception** des expressions vocales, sur les processus de **décodage** de l'émotion à partir des expressions vocales. Ces études s'intéressent principalement à mettre en évidence la capacité des individus à reconnaître/discriminer différentes émotions dans des expressions vocales en l'absence d'indices verbaux ou contextuels. Les études qui proposent des descriptions des caractéristiques vocales impliquées dans les processus de décodage en associant les caractéristiques vocales des expressions aux attributions émotionnelles sont beaucoup plus rares.

Les études qui considèrent simultanément l'encodage (la production) et le décodage (la perception) sont pratiquement inexistantes. La section ci-dessous présente en conséquence les méthodes utilisées et les résultats obtenus par les études de production, dans un premier temps, et par les études de perception, dans un second temps.

1.2.1 Encodage – caractéristiques vocales des émotions exprimées

La communication des émotions par la voix n'est en principe possible que si un ensemble de caractéristiques vocales spécifiques correspond à chaque émotion exprimée. En conséquence, un grand nombre de travaux ont essayé d'identifier les caractéristiques vocales correspondant, le plus souvent, à un nombre restreint d'émotions.

1.2.1.1 La méthodologie

Les études centrées sur l'encodage des expressions vocales émotionnelles se distinguent sur plusieurs aspects méthodologiques: Premièrement, la nature des expressions vocales utilisées et les conditions dans lesquelles elles sont obtenues sont variables. Deuxièmement, les états émotionnels considérés, leur nombre et leur définition diffèrent également d'une étude à l'autre. Troisièmement, les caractéristiques vocales examinées sont également variables. Ces trois points sont successivement développés ci-dessous et suivis de quelques éléments relatifs à l'étude des processus impliqués dans la production des expressions vocales.

Nature et origine des expressions vocales étudiées

Les expressions vocales étudiées sont parfois enregistrées dans des *situations naturellement inductrices* d'émotions. Des situations extrêmement différentes ont été exploitées dans différentes études: "talk-shows" télévisés (e.g. Chung, 2000), séances de psychothérapie (e.g. Eldred & Price, 1958) ou encore communications radio lors d'accidents d'aviation (e.g. Williams & Stevens, 1969).

Ces expressions possèdent une très bonne validité écologique, mais comportent également un certain nombre d'inconvénients: Elles sont produites par un très petit nombre de locuteurs (souvent un seul). Elles incluent un nombre réduit d'états émotionnels pour un même locuteur (parfois un seul). Les états émotionnels vécus par les locuteurs au moment de la production des expressions sont définis a posteriori par les chercheurs, sur la base de critères nécessairement arbitraires. Le contenu verbal des expressions ne peut être contrôlé. Finalement, les conditions d'enregistrements sont souvent mauvaises; les enregistrements comportent des sons/bruits et leur qualité acoustique est généralement réduite.

Une alternative consiste à *induire* des états émotionnels en plaçant les locuteurs *dans des situations contrôlées* en laboratoire (e.g. Bachorowski & Owren, 1995; Sobin & Alpert, 1999). Plusieurs locuteurs peuvent ainsi être placés dans les mêmes conditions (destinées à modifier leur état émotionnel). Différents types d'inductions peuvent être utilisés avec les mêmes locuteurs de manière à éliminer la confusion entre le locuteur et l'état émotionnel induit. Ce cadre permet également d'utiliser le paradigme du contenu verbal constant (standard content paradigm), les locuteurs pouvant être amenés à prononcer une même phrase dans différentes conditions émotionnelles. Les expressions produites dans les différentes conditions sont dès lors réellement comparables, au sens où seule l'expression non-verbale les distingue. Ce cadre de production des expressions vocales comporte pourtant également des inconvénients: En particulier, certains états émotionnels ne peuvent être aisément induits en laboratoire; les contraintes éthiques, d'une part, et la faible implication personnelle des locuteurs dans les situations de laboratoire, d'autre part, ne permettent notamment pas d'induire des émotions très intenses. L'induction se limite donc le plus souvent à manipuler le contexte de manière à produire différents niveaux de stress, souvent assez légers. De plus, les situations identiques dans lesquelles les différents locuteurs sont placés peuvent induire des réactions parfois très différentes d'un locuteur à l'autre. En conséquence, les expressions vocales produites par différents locuteurs dans une même condition ne reflètent pas nécessairement le même état émotionnel et, dans tous les cas, les expressions vocales ne sont que faiblement modifiées par les changements d'états relativement subtils induits.

Une troisième alternative, souvent privilégiée par les auteurs des recherches effectuées dans le domaine des expressions vocales émotionnelles, consiste à utiliser des expressions émotionnelles simulées par des acteurs (e.g. Banse & Scherer, 1996; Davitz, 1964; Fairbanks & Pronovost, 1939; Williams & Stevens, 1972). Contrairement aux expressions enregistrées dans un contexte naturel, les expressions simulées par des acteurs présentent l'avantage de fournir des expressions avec un contenu linguistique constant et correspondant à plusieurs états émotionnels différents pour les mêmes individus. Elles sont, d'autre part, beaucoup plus prononcées que les expressions

enregistrées en situation d'induction émotionnelle qui reflètent en général des modifications très légères de l'état affectif des individus. Le principal reproche formulé à l'encontre des expressions simulées par des acteurs concerne leur manque, supposé, de validité écologique. Dans ce type d'expressions, la composante de communication volontaire est exagérée, alors que la composante d'expressivité "pure/spontanée" – qui correspondrait à des modifications physiologiques associées à un état émotionnel "réel" – est diminuée, voir absente (v. discussion théorique à ce propos, section 1.1.2). En réalité, toutes les expressions comprennent une part de régulation ou de contrôle. Les émotions exprimées dans un "talk-show" ou les expressions induites en laboratoire ne sont notamment pas exemptes de régulations. Il importe surtout de distinguer explicitement ces différents types d'expressions et de garder à l'esprit que les codes de communication sociale sont certainement exagérés dans les expressions "prototypiques" produites par des acteurs.

Selon Juslin & Laukka (2003), seules 7% des 104 études qu'ils ont examinées ont utilisé des procédures d'induction pour obtenir les expressions vocales qu'elles ont analysées; 12% des études ont utilisé des expressions enregistrées dans des contextes naturels et 87% des expressions produites par des acteurs. La disproportion entre le nombre d'études consacrées aux expressions simulées relativement au nombre d'études consacrées aux expressions "naturelles" et induites est manifeste. Les résultats actuellement connus se rapportent donc essentiellement à des expressions "prototypiques" produites par des acteurs.

Etats émotionnels exprimés

La grande majorité des études réalisées à ce jour ont utilisé des catégories émotionnelles, le plus souvent identifiées par des labels correspondant à des émotions dites "fondamentales". Les 104 études examinées dans la revue de Juslin & Laukka (2003) utilisent des catégories émotionnelles; le nombre de catégories utilisées dans ces études varie entre 1 et 15, en moyenne une étude inclut 4 à 5 catégories d'émotions. Les catégories les plus fréquemment étudiées sont: la 'colère', la 'peur', la 'joie' et la 'tristesse'. Les études qui ont explicitement tenté de tester l'existence de différences entre des expressions correspondant à différents types de 'colère', de 'peur', de 'joie' ou de 'tristesse' sont relativement rares. Banse & Scherer (1996) ont introduit une différence entre la colère froide et chaude, l'anxiété et la peur panique, la joie calme et la joie intense, la tristesse déprimée et le désespoir; dans ce cas, la distinction réalisée à l'intérieur de chaque catégorie/famille émotionnelle reflète un niveau d'activation faible versus fort. Burkhardt (2001) a repris en partie cette distinction basée sur le niveau d'activation associé à la catégorie émotionnelle: il distingue la tristesse calme et le désespoir, la colère chaude et froide. On trouve également une distinction entre une joie calme et une joie excitée chez Katz (1998). Juslin & Laukka (2001) ont introduit une distinction liée à

l'intensité des émotions exprimées; leur étude différencie des expressions correspondant à une intensité faible et une intensité forte de colère, de peur, de tristesse, de joie et également de dégoût. Frick (1986) distingue quant à lui des expressions correspondant à une colère frustrée et à une colère agressive.

Parmi les exemples cités ci-dessus, on relèvera que certains auteurs ont inclus des états émotionnels définis à la fois par une famille/catégorie émotionnelle et par une dimension (activation ou intensité). Les états émotionnels sous-jacents aux expressions étudiées sont plus rarement définis uniquement par des dimensions. Quelques exceptions existent toutefois, Murray, Arnott, & Rohwer (1996) se sont par exemple centrés uniquement sur l'activation sous-jacente aux expressions vocales et Fulchner (1991) ainsi que Safer & Leventhal (1977) uniquement sur la valence. L'utilisation d'une dimension unique est évidemment réductrice. A notre connaissance, il n'existe que deux études qui ont tenté d'inclure plusieurs dimensions pour définir les états émotionnels et les expressions qu'elles ont étudiées: une étude de Johnstone (2001, v. aussi Johnstone, Van Reekum, & Scherer, 2001) et une étude de Laukka, Juslin, & Bresin (soumis). Quelques aspects relatifs à l'approche utilisée par Johnstone sont développés ci-dessous, dans la section consacrée à l'étude des processus de production.

Caractéristiques vocales examinées

En ce qui concerne les caractéristiques vocales examinées deux types de méthodes ont été employés par différents auteurs: dans un petit nombre d'études, des *jugements perceptifs* ont été obtenus pour un ensemble de caractéristiques vocales; dans la plupart des études toutefois, des *analyses acoustiques* ont été effectuées.

Les études qui ont produit des évaluations perceptives de caractéristiques vocales sont relativement peu nombreuses. La revue effectuée par Juslin & Laukka (2003) permet d'estimer la proportion des études qui ont utilisés cette approche à environ 8% (6 études ayant utilisé des jugements perceptifs contre 71 études ayant utilisé des mesures acoustiques). Les jugements sont parfois produits par des experts. Dans une étude de van Bezooijen (1984), par exemple, l'auteur et 5 autres spécialistes en phonétique et en linguistique ont effectué un entraînement collectif à l'évaluation de 13 caractéristiques vocales dérivées du système de production de Laver (1980): le degré d'arrondissement et le degré d'extension des lèvres, le degré de tension du larynx, le degré de relâchement du larynx, le degré de "grincement" (creak), de tremblement, de chuchotement, la qualité rauque, la hauteur, l'étendue de la hauteur (pitch range), l'intensité (loudness), le tempo et la précision de l'articulation. D'autres études ont fait appel à des jugements produits par des auditeurs sans expertise particulière et sans entraînement préalable. Dans une étude de Davitz (1964), 20

participants ont ainsi évalué 4 caractéristiques vocales: l'intensité, la hauteur, le timbre et le débit de parole. On relèvera que les caractéristiques vocales qui peuvent être évaluées par des auditeurs sans expertise sont évidemment moins nombreuses et moins spécifiques que les caractéristiques qui peuvent être évaluées par des experts.

Une méthode plus objective pour analyser les caractéristiques vocales des expressions émotionnelles consiste à effectuer des analyses acoustiques. Dans ce domaine également les paramètres extraits varient en fonction des études. Les mesures les plus couramment effectuées correspondent à des paramètres dérivés du contour de la fréquence fondamentale, du contour de l'intensité et de la durée des expressions.

La fréquence fondamentale (F0), exprimée en hertz (Hz), correspond au nombre de répétitions de la période fondamentale du signal acoustique par seconde. Le contour de F0 représente l'évolution de la F0 au cours d'une expression. L'intensité acoustique, exprimée en décibels (db), est dérivée de l'amplitude du signal acoustique. Le contour d'intensité correspond à l'évolution de l'intensité acoustique au cours d'une expression. Les mesures extraites du contour d'intensité et de contour de F0 sont en général descriptives des tendances centrales (moyennes ou médianes) et de la variabilité globale (écarts-types ou écarts entre minima et maxima) des contours. Des indications relatives à la forme des contours de F0 ou d'intensité sont rarement rapportées. Récemment toutefois, Juslin & Laukka (2001) ont examiné la direction globale des contours de F0 (montants versus descendants) et l'attaque des contours d'intensité (degré d'accroissement de l'intensité/amplitude par unité de temps). Mozziconacci (1998) a par ailleurs présenté une description qualitative de la forme des contours de F0 qu'elle a observé pour différents types d'émotions exprimées.

Différents aspects de la durée des expressions ont été rapportés dans différentes études. Les deux mesures les plus courantes sont le débit de parole – par exemple le nombre de syllabes prononcées par minute (ou plus simplement la durée totale des expressions lorsque le contenu verbal est constant) – et la durée proportionnelle des pauses relativement à la durée totale des expressions. Lorsque les expressions sont suffisamment longues, le nombre des pauses est parfois mesuré. Quelques auteurs (e.g. Banse & Scherer, 1996) ont également rapporté la durée relative des parties voisées et des parties non-voisées.

Afin de parvenir à mieux caractériser les expressions correspondant à différentes émotions, quelques auteurs ont récemment tenté de diversifier et d'augmenter le nombre de paramètres acoustiques analysés. Banse & Scherer (1996) ont notamment inclus un ensemble de mesures dérivées du spectre moyen à long terme (long term average spectrum). Dans ce type d'approche, un spectre moyen est calculé pour chaque expression vocale par une analyse de Fourrier inverse qui

permet de décomposer le signal acoustique en composantes fréquentielles associées à différents niveaux d'énergie. L'énergie relative dans différentes bandes de fréquence (par exemple l'énergie comprise entre 0 et 1000 Hz relativement à l'énergie totale) peut ainsi être examinée. Banse & Scherer ont évalué l'énergie relative comprise dans plusieurs bandes de fréquences séparément pour les segments voisés et les segments non-voisé des expressions.

Juslin & Laukka (2001) ont encore inclus d'autres mesures, notamment les perturbations à court terme de la F0 ("jitter") qui correspondent à des fluctuations rapides et aléatoires de la durée d'ouverture/fermeture des cordes vocales de cycle en cycle. Cette mesure a également été utilisée par d'autres auteurs (e.g. Bachorowski & Owren, 1995) qui l'ont présentée comme une mesure indicative de la présence d'un état de stress psychologique chez le locuteur. La hauteur moyenne des formants et la largeur de la bande de fréquence qui contient l'énergie associée aux formants a également été évaluée par Juslin & Laukka (2001). Les formants sont définis comme des régions du spectre acoustique dans lesquelles l'énergie est particulièrement élevée, reflétant les résonances produites par la forme du tractus vocal. Les deux premiers formants définissent avant tout la valeur (catégorie perçue) de la voyelle prononcée; les valeurs des formants sont toutefois également associées à la qualité vocale perçue. Juslin & Laukka (2001) ont également évalué la "précision de l'articulation" définie comme la distance entre les valeurs mesurées des formants et des valeurs de formant neutres (correspondant à la voyelle "schwa"). Plus les valeurs des formants s'approchent de la référence neutre, plus l'effort d'articulation est considéré comme faible.

Relativement aux mesures plus classiques de la F0, de l'intensité ou de la durée des expressions, ces paramètres – "jitter" ou valeurs associées aux formants – sont conçues comme des mesures qui devraient permettre d'améliorer la différenciation entre différentes émotions exprimées. Il s'agit toutefois de mesures qui ne peuvent être évaluées de manière fiable qu'en présence de voyelles soutenues; en d'autres termes, elles requièrent une bonne "qualité phonétique" des enregistrements. Or les émotions tendent à affecter (dégrader) la "qualité phonétique" des expressions verbales. La possibilité d'obtenir des mesures fiables de "jitter" ou de formants dans la parole émotionnelle doit donc encore faire l'objet de plus amples investigations.

Etude des processus de production

Les processus de production – la manière dont l'état émotionnel affecte les expressions vocales – ne sont presque jamais pris en compte dans les études effectuées. A notre connaissance, il n'existe qu'un seul modèle théorique relatif à cette question. Il s'agit du modèle formulé par (Scherer, 1986, 2003) qui postule que différentes dimensions de l'évaluation cognitive antécédente aux réactions

émotionnelles affectent systématiquement les réactions physiologiques qui à leur tour conditionnent les expressions vocales (v. sections théoriques 1.1.1 et 1.1.3).

Une partie des prédictions formulées dans le cadre de ce modèle ont été testées par Johnstone (Johnstone, 2001; Johnstone et al., 2001) Dans une situation de laboratoire (interactions avec un jeu d'ordinateur), Johnstone a manipulé deux dimensions de l'évaluation cognitive: (1) l'implication de l'événement relativement aux buts de l'individu (l'événement favorise ou au contraire entrave les buts de l'individu) et (2) la contrôlabilité de la situation (le contrôle de l'individu sur la situation est soit élevé soit réduit). L'enregistrement des réactions physiologiques des participants a permis de mettre en évidence un effet d'interaction des deux dimensions manipulées. Les situations qui entravaient les buts des participants ont produit une activation plus importante du système nerveux autonome que les situations qui favorisaient leurs buts; mais ce sont les situations qui à la fois entravaient les buts des participants et étaient peu contrôlables qui ont eu l'effet le plus massif sur l'activation autonome (évaluée par le niveau de conductance de la peau). Johnstone (2001) a pu mettre en évidence que dans ces situations défavorables et peu contrôlables la période d'ouverture de la glotte (mesurée par électroglottographie) diminue en association avec l'augmentation de la conductance de la peau. Sur le plan acoustique, les résultats de Johnstone montrent que cela se traduit par une élévation des valeurs de plancher de la fréquence fondamentale ('F0 floor').

Cette étude illustre la possibilité d'intégrer des résultats acoustiques à un modèle théorique des processus de production. L'avantage de ce type d'approche réside notamment dans la définition plus explicite de l'état émotionnel associé aux modifications acoustiques observées. Dans l'exemple présenté, l'élévation des valeurs de plancher de la fréquence fondamentale est associée à l'augmentation de l'activation sympathique des participants qui est elle-même due à une situation perçue comme défavorable et peu contrôlable. Cette définition est plus spécifique qu'une définition qui se réduirait à une catégorie émotionnelle telle que 'stress' ou 'anxiété'.

1.2.1.2 Les principaux résultats

Nous avons relevé ci-dessus que différentes études de production peuvent inclure: différents types d'expressions (simulées, induites ou spontanées) et différents types d'émotions. De plus les mêmes catégories émotionnelles ('peur', 'colère', etc.) peuvent recouvrir des états émotionnels variables dans différentes études. Dans ce contexte, il est évidemment difficile de comparer et de synthétiser les résultats obtenus par les études de production. Nous avons choisi de reproduire ci-dessous la synthèse des résultats proposée par Juslin & Laukka (2003), ainsi que la synthèse des résultats proposée par Scherer (2003). Quelques commentaires et conclusions relativement à ces descriptions des résultats obtenus par les études de production seront ensuite formulés.

Le tableau 1 résume la revue des résultats présentés par Juslin & Laukka (2003). Ces auteurs ont inclus les expressions de colère, de peur, de joie, de tristesse et de tendresse dans leur revue. Ils ont inclus, d'autre part, 7 paramètres acoustiques relativement fréquemment étudiés (partie supérieure du tableau 1); ainsi que 7 paramètres plus rarement utilisés (partie inférieure du tableau 1). Juslin & Laukka ont reclassé les résultats obtenus dans différentes études en 3 niveaux pour certains paramètres acoustiques – par exemple: intensité forte, moyenne ou faible – ou 2 niveaux pour d'autres paramètres – par exemple: contour de F0 montant versus descendant. Dans le tableau 2, nous indiquons pour chaque paramètre acoustique et chaque émotion considérée le niveau le plus fréquemment obtenu dans les études examinées par Juslin & Laukka. Les nombres entre parenthèses indiquent le nombre d'études ayant obtenu ce résultat (1er nombre) sur le nombre total d'études (2ème nombre) ayant rapporté un résultat pour cette émotion et ce paramètre dans la revue effectuée par Juslin & Laukka.

Tableau 1: synthèse des résultats tirée de la revue de Juslin et Laukka (2003, pp. 792-799)

Paramètre	colère	peur	joie	tristesse	tendresse
Intensité (moy) (forte - moyenne - faible)	forte (30/32)	forte (11/22)	forte (20/26)	faible (29/32)	faible (4/4)
Intensité variabilité (forte - moyenne - faible)	forte (9/12)	forte (7/12)	forte (8/13)	faible (8/11)	
F0 (moy) (haute - moyenne - basse)	haute (33/43)	haute (28/39)	haute (34/38)	basse (40/45)	basse (4/5)
F0 variabilité (forte - moyenne - faible)	forte (27/35)	faible (17/32)	forte (33/36)	faible (31/34)	faible (5/5)
F0 contours (montant - descendant)	montant (6/8)	montant (6/6)	montant (7/7)	descendant (11/11)	descendant (3/4)
Energie hautes fréqu. (forte - moyenne - faible)	forte (22/22)	forte (8/16)	forte (13/17)	faible (19/19)	faible (3/3)
Débit de parole (rapide - moyen - lent)	rapide (28/35)	rapide (24/29)	rapide (22/33)	lent (30/36)	lent (3/4)
Régularités microstruct.[b] (régulier - irrégulier)	irrégulier (3/3)	irrégulier (2/2)	régulier (2/2)	irrégulier (4/4)	régulier (1/1)
Proportion de pauses (forte - moyenne - faible)	faible (8/8)	faible (4/9)	faible (3/6)	forte (11/12)	forte (1/1)
Précision articulation (haute - moyenne - basse)	haute (7/7)	?[a] (2 - 2 - 2)	haute (3/5)	basse (6/6)	basse (1/1)
Formant 1 (hauteur) (haut - moyenne - bas)	haut (6/6)	bas (3/4)	haut (5/6)	bas (5/6)	
Formant 1 (largeur) (étroit - large)	étroit (4/4)	large (2/2)	étroit (2/3)	large (3/3)	
"Jitter" (fort - faible)	fort (6/7)	?[a] (4 - 4)	fort (5/8)	faible (5/6)	
"Glottal waveform"[c] (abrupte - arrondie)	abrupte (6/6)	arrondie (4/6)	abrupte (2/2)	arrondie (4/4)	

[a] Un nombre égal d'études ont rapporté différents niveaux pour ce paramètre et cette émotion (le nombre d'études ayant obtenu chaque niveau considéré pour ce paramètre est représenté entre parenthèses)

[b] Les irrégularités microstructurales sont définies comme des irrégularités à court terme au niveau de la F0, de l'intensité et/ou de la durée. L'irrégularité est théoriquement associée aux expressions émotionnelles négatives.

[c] La forme du signal de la source vocale peut être obtenu par filtrage inverse (Laukkanen, Vilkman, Alku, & Oksanen, 1996).

Les proportions rapportées entre parenthèses dans le tableau 1 révèlent, qu'à partir du moment où un nombre relativement important d'études ont examiné un paramètre acoustique pour une émotion donnée, l'unanimité des différentes études concernant la valeur de ce paramètre pour cette émotion est assez rare. De tels cas existent néanmoins; ainsi les 22 études ayant examiné la proportion d'énergie dans les hautes fréquences pour les expressions de colère rapportent une forte proportion d'énergie dans les hautes fréquences pour ces expressions. Le désaccord entre les différentes études et parfois assez important: seules 17 études sur 32 études examinées par Juslin & Laukka rapportent par exemple que les expressions de peur sont associées à une faible variabilité de la F0, 6 études ont rapporté une variabilité moyenne de la F0 et 9 études une variabilité forte de la F0 pour les expressions de peur. Cette variabilité des résultats est probablement en grande partie attribuable au problème, déjà plusieurs fois évoqué, de la variabilité de la définition des états émotionnels regroupés sous le même label. Les expressions de 'peur' pourraient par exemple correspondre à des expressions 'd'inquiétude' dans certaines études et à des expressions de 'peur panique' dans d'autres études.

Les paramètres acoustiques fréquemment utilisés, représentés dans la partie supérieure du tableau 1, ne paraissent pas en mesure de différencier les 7 catégories émotionnelles examinées. A ce niveau de représentation, ces paramètres semblent tous effectuer la même distinction entre les expressions de colère, de peur et de joie, d'une part, et les expressions de tristesse et de tendresse, d'autre part. La seule exception concerne la variabilité de la F0 qui est plus souvent rapportée comme faible pour les expressions de peur[1] alors qu'elle est plus fréquemment rapportée comme forte pour les expressions de colère et de joie. Les paramètres moins fréquemment utilisés, représentés dans la partie inférieure du tableau, présentent des patterns différents pour les 5 catégories émotionnelles considérées. Les résultats pour ces paramètres sont toutefois peu nombreux et les résultats obtenus par différentes études ne sont pas toujours identiques. D'avantage de résultats concernant ces paramètres sont donc nécessaires avant de pouvoir en dériver des conclusions définitives.

Le tableau 2 est reproduit de la revue de Scherer (2003, p. 233). Les flèches dirigées vers le haut désignent un accroissement du paramètre acoustique considéré pour l'expression émotionnelle envisagée; les flèches dirigées vers le bas correspondent à une diminution de la valeur du paramètre (ou des contours de F0 descendants). La synthèse de Scherer inclut 6 catégories émotionnelles: le stress, la colère/rage, la peur/panique, la tristesse, la joie/intense et l'ennui. Les spécifications des catégories émotionnelles proposées après les barres obliques dérivent de l'idée que les expressions

[1] Le désaccord en ce qui concerne la variabilité de la F0 pour les expressions de peur est toutefois assez important (cf. commentaire ci-dessus).

de colère, de peur et de joie habituellement étudiées correspondraient à des versions actives de ces émotions, c'est-à-dire la colère chaude, la peur panique et la joie intense. Cette revue inclut d'autre part 7 paramètres acoustiques qui ont été fréquemment évalués pour les expressions émotionnelles considérées (ou au moins pour une partie d'entre elles): l'intensité acoustique moyenne, la F0 moyenne (ou une valeur minimale de la F0), la variabilité de la F0, l'étendue de la F0, la forme des contours de F0, la proportion d'énergie dans les hautes fréquences et le débit de parole.

Tableau 2: synthèse des résultats obtenus par les études de production selon Scherer (2003, p. 233)

Synthetic compilation of the review of empirical data on acoustic patterning of basic emotions (based on Johnstone and Scherer, 2000)

	Stress	Anger/rage	Fear/panic	Sadness	Joy/elation	Boredom
Intensity	↗	↗	↗	↘	↗	
F0 floor/mean	↗	↗	↗	↘	↗	
F0 variability		↗		↘	↗	↘
F0 range		↗	↗(↘)	↘	↗	↘
Sentence contours		↘		↘		
High frequency energy		↗	↗	↘	(↗)	
Speech and articulation rate		↗	↗	↘	(↗)	↘

Les critères d'inclusion/exclusion d'un paramètre acoustique ou d'une catégorie d'expression émotionnelle dans une revue de la littérature sont toujours plus ou moins arbitraires. La comparaison des revues de Scherer (2003) et de Juslin & Laukka (2003) permet de constater que ces auteurs ont notamment inclus des états émotionnels en partie différents – 'stress' et 'ennui' dans la revue de Scherer et 'tendresse' dans la revue de Juslin et Laukka – et différents paramètres acoustiques – Scherer n'inclut notamment pas les paramètres peu fréquents, alors que Juslin & Laukka ont inclus l'étendue de F0 (F0 range) dans la catégorie "variabilité de la F0". Les résultats qui recouvrent les mêmes émotions et les même paramètres acoustiques sont toutefois identiques dans les deux revues; à l'exception des contours de F0 qui seraient descendants pour la colère selon la revue proposée par Scherer, alors que Juslin & Laukka indiquent que 6 des 8 études qu'ils ont considérées rapportent de contours montants pour les expressions de colère.

Ces deux revues permettent de tirer des conclusions qui ont été également formulées dans d'autres publications (par exemple Davitz, 1964; Frick, 1985; Johnstone & Scherer, 2000; Scherer, 1986). Si l'on accepte le postulat selon lequel les expressions de joie, de colère et de peur habituellement étudiées correspondent à des états émotionnels fortement activés (joie intense, colère chaude, peur panique), les résultats issus des études de production indiquent que les paramètres acoustiques habituellement mesurés reflètent essentiellement l'*activation émotionnelle*. Les émotions qui incluent une activation forte – telle que la colère, la peur panique, la joie intense – présentent un accroissement des valeurs de F0 et d'intensité et une diminution de la durée de différents segments correspondant à une accélération de la parole. Alors que les états qui incluent un degré d'activation faible – tels que la tristesse, l'ennui ou la tendresse – présentent une diminution des valeurs de F0 et

d'intensité, ainsi qu'une augmentation de la durée de différents segments. Un constat très similaire a été effectué par Sundberg, Iwarsson, & Hagegård (1995) dans le contexte d'une étude de l'expression émotionnelle dans la voix chantée. Ces auteurs ont observé que des extraits de musique classique interprétés de manière expressive (émotionnelle) par un chanteur professionnel sont réalisés avec un tempo ralenti et une intensité diminuée pour les extraits qu'ils ont identifiés comme "non-agités" ; alors que le tempo est rapide et l'intensité élevée dans les interprétations émotionnelles des extraits qui possèdent un caractère musical "agité".

Des études isolées présentent parfois des patterns acoustiques spécifiques pour différentes catégories/familles d'émotions exprimées. Mais ces profils acoustiques sont rarement répliqués d'une étude à l'autre. Sur le plan général, le consensus semble donc bien se limiter à un effet de l'activation sous-jacente aux expressions étudiées. Deux types d'explications, au moins, ont été avancées à ce sujet. La première concerne le problème, déjà plusieurs fois évoqué, de la définition des états émotionnels sous-jacents aux expressions étudiées. Des profils différenciés pour différentes catégories émotionnelles ne peuvent être consensuels si ces catégories recouvrent en réalité différents états émotionnels (et donc différentes expressions) dans différentes études. La seconde explication a trait aux paramètres acoustiques utilisés pour définir les profils acoustiques des expressions émotionnelles. Les paramètres mesurés reflèteraient essentiellement la dimension d'activation émotionnelle et l'utilisation d'autres paramètres – mieux choisis - permettrait une meilleure différenciation des différents états émotionnels sur le plan acoustique. Les résultats résumés par Juslin & Laukka (2003, v. tableau 1) tendent à renforcer cette hypothèse. Ils indiquent que les paramètres acoustiques moins fréquemment utilisés pourraient éventuellement contribuer à différencier les expressions émotionnelles examinées.

1.2.2 Décodage – reconnaissance des expressions vocales émotionnelles

Les études de la perception des expressions vocales visent en premier lieu à évaluer la possibilité pour des groupes d'auditeurs de reconnaître (discriminer) des expressions correspondant à différentes catégories émotionnelles. Quelques études se sont par ailleurs intéressées aux caractéristiques vocales qui permettent aux auditeurs de reconnaître les expressions vocales émotionnelles. Les méthodes utilisées par ces deux types d'études et les principaux résultats qu'elles ont obtenus sont présentés ci-dessous.

1.2.2.1 La méthodologie

Les principales méthodes utilisées par les études qui ont tenté d'évaluer la possibilité de reconnaître différents types d'émotions exprimées seront présentées et discutées dans un premier temps. Les

méthodes utilisées par les études qui se sont penchées sur les caractéristiques vocales qui interviennent dans le processus de reconnaissance seront présentées ensuite.

Reconnaissance/discrimination des catégories émotionnelles exprimées

Les différents types d'expressions décrits ci-dessus, dans la section consacrée aux méthodes utilisées par les études de production, ont été exploités également dans des études de reconnaissance des émotions exprimées. La majorité des études de perception ont toutefois été réalisées en utilisant des expressions émotionnelles simulées par des acteurs et correspondant à un nombre relativement réduit de catégories émotionnelles. Dans ce domaine d'étude, le contenu verbal des expressions ne doit, en principe, pas porter de significations émotionnelles, de manière à ne pas influencer les attributions des auditeurs. Le contenu des expressions correspondant à différentes émotions est donc en général constant et "neutre", au sens ou les locuteurs/acteurs prononcent des phrases avec une signification non-émotionnelle ou, parfois, des séquences de syllabes sans signification (e.g. Banse & Scherer, 1996).

Différentes méthodes ont été utilisées dans ce domaine pour recueillir les jugements des auditeurs relativement aux émotions exprimées. Le plus souvent, des expressions vocales correspondant à des catégories émotionnelles – telles que la joie, la peur, la tristesse ou la colère – sont présentées à des auditeurs qui ont pour tâche de choisir, parmi une liste de catégories émotionnelles, la catégorie qui correspond à l'expression vocale présentée. La liste des catégories disponibles correspond en général aux catégories émotionnelles qui sont théoriquement exprimées dans les expressions vocales. Dans cette procédure de choix forcé, la tâche des auditeurs n'est donc pas de *reconnaître* les émotions exprimées mais de les *discriminer* en fonction des catégories qui leur sont soumises.

Parallèlement au nombre d'émotions théoriquement exprimées dans les enregistrements étudiés, le nombre d'alternatives de réponse varie d'étude en étude: 14 catégories émotionnelles ont été, par exemple, proposées aux auditeurs dans l'étude de Banse & Scherer (1996); alors qu'on trouve 7 catégories (incluant une catégorie "neutre") chez Mozziconacci (1998). Beaucoup d'études n'incluent toutefois pas plus de 5 à 6 catégories différentes (e.g. Breitenstein, Van Lancker, & Daum, 2001; Scherer, Banse, & Wallbott, 2001). Le nombre d'expressions correctement reconnues est évidemment en partie fonction du nombre de catégories émotionnelles exprimées et du nombre d'alternatives de réponse différentes. Dans une étude qui utiliserait 5 alternatives de réponse, un auditeur a 20% de chance de donner une réponse correcte au hasard; alors que la probabilité de répondre correctement au hasard n'est que de 10% dans une étude qui utiliserait 10 alternatives de réponse. La proportion de réponses correctes pour une catégorie émotionnelle spécifique est d'autre part fonction des biais de réponse qui peuvent éventuellement apparaître dans ces études. Si une

catégorie émotionnelle est systématiquement choisie plus fréquemment que les autres (alors que toutes les catégories sont exprimées avec la même fréquence), la probabilité que les expressions qui correspondent à cette catégorie soient reconnues correctement est plus élevée que pour les catégories qui sont choisies moins fréquemment.

Par ailleurs, la procédure basée sur la discrimination peut favoriser des taux de reconnaissance élevés dans les cas où certaines catégories peuvent être facilement discriminées de l'ensemble des autres catégories. Dans l'hypothèse où certaines propriétés dimensionnelles des expressions vocales – telles que la valence ou l'activation – pourraient être facilement identifiées, la présence d'une catégorie opposée aux autres catégories sur une telle dimension faciliterait notamment sa discrimination (Russell, 1994). Ainsi une étude qui inclurait les réponses alternatives 'joie', 'peur', 'colère' et 'tristesse' pourrait obtenir des taux de reconnaissance très élevés, basés uniquement sur la discrimination des expressions positives versus négatives (qui permettrait d'identifier les expressions de 'joie') et sur la discrimination des expressions faiblement activées versus fortement activées (qui permettrait d'identifier les expressions de tristesse).

D'autres procédures ont également, bien que plus rarement, été utilisées pour évaluer les attributions émotionnelles relatives aux expressions vocales. Certains auteurs ont par exemple donné la possibilité aux auditeurs de sélectionner plus d'une catégorie pour chaque expression vocale, de manière à indiquer la présence d'un "mélange" entre plusieurs catégories émotionnelles simultanément exprimées (e.g. Mozziconacci, 1998; Scherer et al., 2001). Ces réponses multiples non-systématiques sont toutefois difficiles à traiter sur le plan statistique.

Une autre possibilité consiste à demander aux auditeurs d'évaluer l'intensité de plusieurs émotions exprimées pour chaque expression vocale présentée (e.g. Frick, 1986). Dans ce cas, les auditeurs indiquent une intensité nulle lorsqu'ils n'identifient pas la catégorie émotionnelle considérée dans l'expression présentée ou un degré d'intensité correspondant à l'intensité de l'émotion qu'ils perçoivent dans cette expression. Cette procédure ne nécessite pas, en principe, qu'une discrimination soit effectuée entre les catégories émotionnelles dont l'intensité est évaluée; elle permettrait donc en théorie de résoudre les problèmes associés à la discrimination des catégories. Toutefois, un très grand nombre d'intensités nulles sont en général rapportées dans ce contexte. Il existe donc un risque de "glissement" sur le plan de l'utilisation des échelles d'intensité. Il ne peut être exclu notamment que les auditeurs sélectionnent (discriminent) une catégorie et lui attribue une intensité relativement élevée tout en attribuant une intensité nulle à l'ensemble des autres catégories. Paradoxalement, cette utilisation des échelles d'intensité est en général "souhaitée" par les

chercheurs qui postulent le plus souvent qu'un seul type d'émotion est exprimé dans chaque expression vocale.

Une dernière possibilité d'évaluer les attributions émotionnelles relativement aux expressions vocales consiste à évaluer les expressions sur une ou plusieurs dimensions sous-jacentes aux émotions exprimées. Cette approche a été utilisée déjà par Davitz (1964) qui a demandé à un groupe d'auditeurs d'évaluer un ensemble d'expressions vocales sur 3 dimensions proposées par Osgood, Suci, & Tannenbaum (1957): la valence, l'activation ("activity") et le contrôle ("strength"). D'autres tentatives de caractériser les expressions vocales émotionnelles en fonction de différentes dimensions perçues ont depuis été effectuées (v. Laukka et al., soumis). Certains auteurs ont notamment combiné une approche dimensionnelle avec une approche catégorielle en ajoutant l'évaluation d'une ou plusieurs dimensions à la sélection d'une catégorie (e.g. Breitenstein et al., 2001).

Caractéristiques vocales impliquées dans les processus de reconnaissance

Dans le cadre de l'étude de la reconnaissance des expressions vocales émotionnelles, un petit nombre de travaux se sont intéressés aux processus de décodage. Ces études ont tenté d'identifier les caractéristiques vocales utilisées par les auditeurs pour former des attributions émotionnelles à partir des expressions vocales. Trois types d'approches ont été utilisées dans ce domaine.

Quelques études ont établi des corrélations multiples entre les caractéristiques acoustiques des expressions vocales et les attributions émotionnelles effectuées par des auditeurs. Cette première approche fournit des indications concernant les caractéristiques susceptibles d'avoir influencé les attributions émotionnelles des auditeurs (Banse & Scherer, 1996; Scherer, 2003; van Bezooijen, 1984).

Une deuxième approche consiste à éliminer (masquer) une partie de l'information contenue dans les expressions. Dans ce domaine la technique la plus fréquemment utilisée consiste à éliminer par filtrage toutes les fréquences qui dépassent un seuil donné (low-pass filtering), ce qui a pour but de supprimer les informations relatives au timbre vocal ainsi que le contenu phonétique des expressions, alors que l'essentiel des aspects rythmiques et mélodiques restent préservés. D'autres techniques – telles que le découpage des expressions en segments courts et leur recomposition dans un ordre aléatoire (randomized splicing) – peuvent être utilisées afin de conserver au contraire le timbre vocal et supprimer les aspects de rythme et de mélodie. Cette approche a permis de démontrer qu'il reste possible d'identifier l'émotion exprimée même lorsque l'on supprime certaines dimensions de l'information. L'émotion serait donc communiquée à la fois par les aspects mélodiques et rythmiques de la voix ainsi que par certains aspects du timbre vocal (v. Scherer,

Feldstein, Bond, & Rosenthal, 1985, pour une discussion des caractéristiques et des résultats de différentes techniques de masquage).

La troisième approche consiste à manipuler certaines caractéristiques des expressions en utilisant des techniques de synthèse ou resynthèse vocale. Cette approche permet la manipulation expérimentale simultanée de plusieurs paramètres vocaux dont les effets directs et les effets d'interactions sur les attributions émotionnelles peuvent être évalués. Scherer & Oshinsky (1977) ont par exemple étudié l'effet de la variation de l'amplitude, du niveau, du contour et de la variabilité de la F0, ainsi que l'effet de la variation du rythme, de la richesse harmonique et de la tonalité sur les attributions émotionnelles. Les techniques de resynthèse permettent d'utiliser des voix naturelles qui sont enregistrées, digitalisées puis reproduites avec des modifications systématiques de certains paramètres. Dans une publication de 1988, Bergmann, Goldbeck, & Scherer ont par exemple manipulé le niveau, la variabilité, l'écart et le contour de F0, l'intensité et la durée de productions vocales réelles. Plus récemment, Mozziconacci (1998) a également manipulé des aspects relatifs aux contours de F0 et à la durée d'un ensemble d'expressions, en tentant de définir les valeurs optimales pour la communication de plusieurs catégories d'émotions.

Les études qui ont à ce jour utilisé ce type d'approche ont confirmé l'intervention de plusieurs dimensions vocales différentes dans le processus d'attribution émotionnelle. Parmi les dimensions manipulées par ces études, l'évolution de la fréquence fondamentale au fil de l'expression (contour de F0) semble jouer un rôle particulièrement important. Le rôle de l'intonation, et en particulier du contour de F0, dans la communication vocale des émotions sera développé de manière plus détaillée dans la section consacrée au rôle de l'intonation dans la communication émotionnelle (section 1.4).

1.2.2.2 Les principaux résultats

Comme pour les études de production, il est assez difficile de comparer et de synthétiser les résultats obtenus par les études consacrées à la reconnaissance (discrimination) des expressions vocales. Les différentes études effectuées dans ce domaine ont utilisé différents types d'expressions et surtout différentes catégories émotionnelles, plus ou moins nombreuses. Des synthèses des résultats sont toutefois assez fréquemment proposées par les principaux auteurs des recherches effectuées dans ce domaine.

Dans une revue de la littérature basée sur environ 30 études effectuées avant le milieu des années 1980, Scherer (1989) indique que le pourcentage de reconnaissance correcte des expressions vocales est d'environ 60%, toutes émotions confondues, dans ces études. Cet auteur estime que ce pourcentage est approximativement 5 fois plus élevé que ce qui serait obtenu si les auditeurs répondaient en choisissant une émotion au hasard. Dans une revue plus récente, Scherer et al.

(2001) rapportent également un pourcentage moyen de reconnaissance correcte de 66% (avec 5 alternatives de réponse) pour 11 études effectuées dans différents pays occidentaux.

Récemment Juslin & Laukka (2003) ont présenté une méta-analyse incluant 39 études ayant utilisé un paradigme de choix forcé. Ils ont traduit les pourcentages de reconnaissance correcte rapportés dans ces études par un index (effect size index, π) proposé par Rosenthal & Rubin (1989) pour la méta-analyse des études de jugements effectuées dans un paradigme de choix forcé. Cet index permet de comparer les résultats d'études incluant différents nombres d'alternatives de réponses en transformant ces résultats pour les exprimer sur une même échelle de "choix dichotomique". Sur une telle échelle, la valeur .50 (50%) correspond à des réponses données au hasard et la valeur 1.00 à des réponses correctes à 100%. Juslin & Laukka (2003) ont inclus 5 catégories émotionnelles qui sont relativement bien reconnues: la colère, la peur, la joie, la tristesse et la tendresse. Dans le tableau 3, nous reproduisons l'index π moyen rapporté par ces auteurs pour chacune de ces 5 catégories émotionnelles et pour la reconnaissance globale des émotions. L'index π le plus faible (minimum) et l'index π le plus élevé (maximum), ainsi que le nombre d'études et le nombre total de locuteurs inclus dans cette méta-analyse sont également reproduit dans le tableau 3.

Tableau 3: Résultats de la méta-analyse publiée par Juslin et Laukka (2003, p. 787)

	colère	peur	joie	tristesse	tendresse	total**
Moyenne index π*	.93	.88	.87	.93	.82	.90
Minimum index π*	.77	.65	.51	.80	.69	.69
Maximum index π*	1.00	1.00	1.00	1.00	.89	1.00
Nombre d'études	32	26	30	31	6	38
Nombre de locuteurs	278	273	253	225	49	473

*Les valeurs rapportées correspondent à la méta-analyse de Juslin et Laukka (2003) pour les études intra-culturelles (les résultats des études qui ont testé la reconnaissance inter-culturelle des expressions vocales ne sont pas inclus)
** La colonne 'total' reflète les taux de reconnaissance émotionnelle rapportés dans 38 études indépendamment des émotions qu'elles ont étudiées. Ce résultat comprend donc implicitement un grand nombre de catégories émotionnelles utilisées dans différentes études (non seulement les catégories représentées dans les colonnes du tableau).

Les résultats de la méta-analyse de Juslin et Laukka corroborent une constatation effectuée également dans d'autres revues de la littérature (e.g. Johnstone & Scherer, 2000; Scherer, 2003) relativement aux pourcentages de reconnaissance obtenus pour différentes catégories d'émotions. Dans la grande majorité des études, la tristesse et la colère sont les catégories émotionnelles les mieux reconnues. Les expressions de peur et de joie sont relativement moins bien reconnues. Ce résultat a été régulièrement répliqué malgré la variabilité des expressions et des méthodes utilisées dans différentes études. Dans le même sens, les expressions de dégoût sont systématiquement très

mal reconnues dans les études qui ont inclus cette catégorie d'expression. Scherer (Johnstone & Scherer, 2000; Scherer, 2003) a mis en évidence que les expressions moins bien reconnues sur le plan vocal – notamment les expressions de dégoût et de joie – sont très bien reconnues sur le plan des expressions faciales. Cette observation permet d'émettre l'hypothèse que certaines expressions – telles que les expressions de joie et de dégoût – seraient "préférentiellement" communiquées par le canal facial/visuel, alors que d'autres expressions – telles que les expressions de colère et de tristesse – seraient préférentiellement communiquées par le canal vocal/auditif.

Les confusions entre les émotions exprimées et les émotions reconnues/discriminées sont assez rarement rapportées dans les études publiées. Plusieurs auteurs relèvent toutefois que les matrices de confusions représentent une source d'information essentielle, notamment relativement aux biais de réponse qui interviennent probablement dans la plupart des études (v. Juslin & Laukka, 2003). Banse & Scherer (1996) ont proposé de considérer les confusions comme des indicateurs de similarité perçue entre les expressions confondues. Dans leur étude publiée en 1996, ces auteurs ont rapporté que les émotions appartenant à une même "famille" – mais correspondant à différents niveaux d'activation comme la "colère chaude" et la "colère froide" – sont plus souvent confondues entre elles qu'avec des émotions appartenant à d'autres "familles". Un autre facteur de confusion (ou une autre dimension de similarité) pourrait être, justement, l'activation. Les expressions 'exaltées' (joie très activée) sont, par exemple, assez souvent confondue avec les expressions de 'désespoir' (tristesse fortement activée). Selon Banse & Scherer (1996), une troisième dimension de similarité serait la valence, les expressions positives sont plus souvent confondues entre elles qu'avec des expressions négatives. Ces hypothèses sont en partie confortées par les résultats obtenus par Green & Cliff (1975). Ces auteurs ont recueilli des jugements de similarité pour un ensemble d'expressions vocales émotionnelles. Une analyse multidimensionnelle leur a permis d'identifier les deux dimensions principales sous-jacentes aux jugements de similarité. Ils ont interprété ces dimensions comme correspondant à la valence et à l'activation sous-jacentes aux émotions exprimées.

Au-delà des confusions (ou des similarités perçues), les études de perception ont largement confirmé que les émotions exprimées par la voix, ou au moins différents aspects des émotions – telles que la valence, l'activation ou la famille émotionnelle – peuvent être correctement reconnues. Cette conclusion contraste avec les résultats obtenus sur le plan acoustique où la synthèse de la littérature indique que les caractéristiques vocales habituellement mesurées semblent refléter essentiellement le niveau d'activation sous-jacent aux expressions émotionnelles. Si des auditeurs sont capables de différencier des expressions vocales correspondant à différentes familles

émotionnelles, des différences acoustiques spécifiques à différents types d'émotions sont nécessairement présentes dans les expressions et devraient pouvoir être mesurées.

Cette dernière constatation doit cependant être modérée, dans la mesure où les résultats relatifs à la reconnaissance des émotions exprimées et les résultats relatifs aux caractéristiques vocales correspondant à différentes émotions exprimées sont le plus souvent obtenus dans des études distinctes. Les études centrées sur la description des caractéristiques vocales correspondant à différentes émotions exprimées rapportent très rarement des résultats détaillés concernant la reconnaissance des émotions exprimées. Lorsque des études de jugement sont effectuées dans ce cadre, elles sont en général destinées à sélectionner les expressions émotionnelles qui sont "le mieux reconnues". Les caractéristiques vocales (paramètres acoustiques) des expressions "bien reconnues" sont ensuite extraites et analysées. La proportion des expressions vocales correctement reconnues par un groupe d'auditeurs est donc parfois rapportée dans les études de la production. En revanche, les confusions effectuées par les auditeurs ne sont en général pas examinées dans ce contexte. Il semble donc particulièrement important d'examiner à la fois la production (les caractéristiques vocales) et la perception (la reconnaissance) des mêmes expressions vocales émotionnelles, notamment afin d'évaluer dans quelle mesure *la communication* peut (ou ne peut pas) être expliquée par les caractéristiques vocales mesurées.

Dans les sections qui suivent, deux aspects de l'étude des expressions vocales émotionnelles seront présentés en plus de détails. La section 1.3 développe une proposition de Scherer (1978, 2003) relativement à la problématique (évoquée dans le paragraphe précédent) de l'étude simultanée de la production et de la perception des expressions vocales émotionnelles. La section 1.4 développe, quant à elle, la question du rôle de l'intonation dans la communication vocale des émotions.

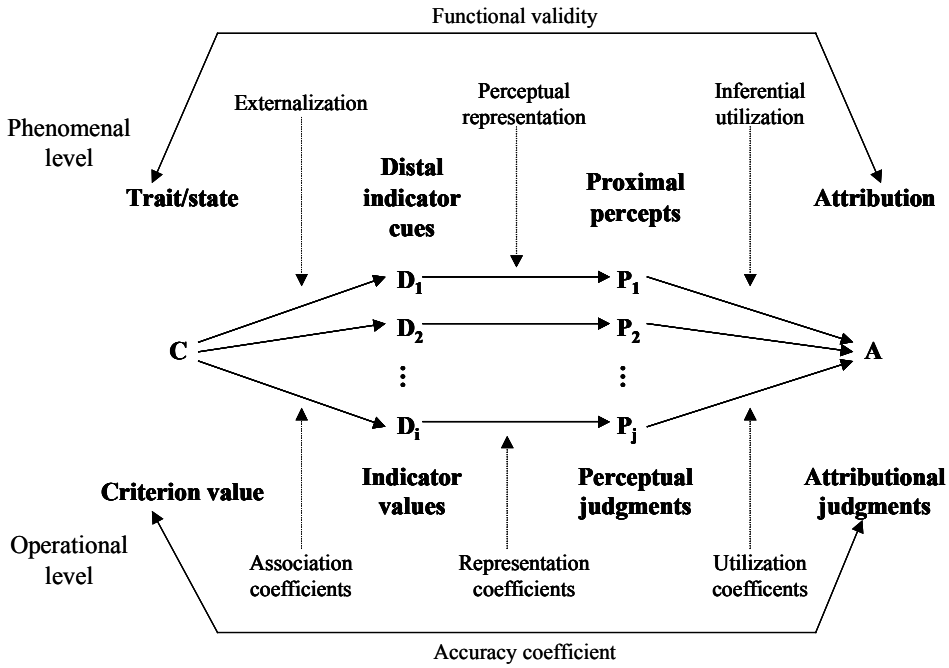
1.3 Paradigme du modèle en lentille de Brunswik

Un paradigme dérivé du modèle en lentille (lens model) de Brunswik est utilisé comme cadre pour la recherche présentée dans cette thèse. La section ci-dessous décrit, dans un premier temps, la proposition de Scherer (1978, 2003) concernant l'utilisation du modèle en lentille de Brunswik pour l'étude de la communication non-verbale. Quelques concepts associés à ce modèle, ainsi que quelques exemples de l'application de ce modèle à différentes problématiques de recherche seront présentés ensuite.

1.3.1 Le modèle en lentille – paradigme pour l'étude de la communication non-verbale

Dans un ouvrage de 1956, Brunswik a formulé un certain nombre de principes théoriques et méthodologiques relatifs à l'étude de la perception. Il a appliqué ces principes à l'étude de différents phénomènes parmi lesquels figurent entre autres les constances et les illusions perceptives mais également l'attribution de traits psychologiques à partir de l'observation de l'aspect extérieur d'un individu. L'approche utilisée par Brunswik a été reprise par Scherer (1978) pour étudier les jugements relatifs à des traits de personnalité sur la base de l'expression vocale. A plusieurs reprises, cet auteur a ensuite proposé d'utiliser une version modifiée du modèle en lentille (Lens Model) de Brunswik comme paradigme pour la recherche sur la communication non-verbale, et en particulier pour l'étude de la communication vocale des émotions (e.g. Scherer, 1982, 2003). Cette adaptation du modèle de Brunswik peut être représentée par la figure 1 (reproduite de Scherer, 1978).

Figure 1: Modèle en lentille proposé par Scherer (1978)



Dans ce modèle, les états internes – dans le cas présent les émotions – sont extériorisés sous la forme d'*indices distaux* (*distal indicators*) qui correspondent dans le contexte de la communication vocale aux caractéristiques acoustiques de la voix. La notion d'*extériorisation* (*externalization*) recouvre à la fois la communication intentionnelle des états internes et les réactions comportementales et physiologiques involontairement produites. Sur le plan opérationnel, les états internes sont représentés par des *valeurs de critère* (*criterion values*) et les indices distaux par des *valeurs d'indicateurs* (*indicator values*). Les indices distaux sont représentés de manière proximale par des *percepts* qui sont le résultat du traitement perceptif réalisé par l'observateur. Sur le plan

opérationnel, les percepts peuvent être évalués par des *jugements* (*perceptual judgments*) exprimés sous forme de scores sur des échelles/dimensions psychophysiques. Les corrélations entre valeurs d'indicateurs et jugements perceptifs sont désignées par le terme de *coefficient de représentation* (*representation coefficient*), elles indiquent le degré de précision de la projection des indices distaux dans l'espace perceptif de l'individu. L'*attribution* d'un état est le résultat de processus d'inférence basés sur la perception des indices distaux. Ces attributions peuvent être évaluées en obtenant à nouveau des jugements de la part d'observateurs mais cette fois sur des dimensions psychologiques. Les corrélations entre jugements perceptifs et attributions sont représentées dans le modèle par les *coefficients d'utilisation* (*utilization coefficients*) qui donnent une mesure de l'utilisation (ou du poids) de chaque indice perçu lors de l'inférence d'un état. L'exactitude des attributions relativement à l'état objectivement observé de l'individu est définie sur le plan opérationnel par la corrélation entre les valeurs de critère et les attributions (*coefficients d'exactitude/accuracy*).

Ce modèle permet de spécifier et de distinguer plusieurs étapes impliquées dans le processus de communication. La partie gauche du modèle correspond aux processus d'encodage de l'émotion dans la voix, alors que la partie droite recouvre les processus de décodage. En opérationnalisant et en mesurant toutes les étapes décrites par le modèle, il devient possible de représenter le processus de communication et d'évaluer l'importance relative de différentes caractéristiques vocales au niveau de l'encodage et du décodage.

Récemment Scherer (2003) a considéré la possibilité d'inclure un nombre plus important d'étapes dans la description du processus de communication (v. aussi Scherer, Johnstone, & Klasmeyer, 2003). Les "variables proximales" (*proximal percepts* dans la figure 1) sont notamment susceptibles d'être définies à plusieurs niveaux: Premièrement, les propriétés des expressions vocales produites par un locuteur peuvent être modifiées (dégradées ou filtrées par le média de communication, couvertes par des bruits ambiants) avant de parvenir à l'émetteur. Deuxièmement, le système perceptif de l'auditeur fonctionne lui-même comme un filtre, il sélectionne et intègre différents aspects du signal acoustique qui lui parviennent. Finalement les caractéristiques acoustiques perçues sont catégorisées/organisées dans le système de représentation de l'auditeur. Dans la proposition originale de Scherer (1978), représentée par la figure 1, seul ce dernier niveau – les représentations/ jugements concernant les caractéristiques vocales des expressions – est représenté. Les autres aspects du processus de communication mériteraient toutefois d'être considérés dans un examen complet du processus d'inférence des émotions exprimées par la voix.

1.3.2 Quelques concepts associés au modèle en lentille de Brunswik

La suite de cette section sera consacrée à introduire quelques aspects de la conceptualisation associée au modèle en lentille de Brunswik. Récemment Hammond & Stewart (2001) ont édité un ouvrage qui reproduit les principaux écrits de Brunswik, ainsi que des commentaires relatifs aux concepts développés dans ces textes. L'objectif des paragraphes qui suivent n'est pas d'exposer la totalité de ces concepts (le lecteur intéressé pourra consulter l'ouvrage de Hammond et Stewart, 2001, à cette fin), mais d'introduire un certain nombre de points qui peuvent être appliqués à la problématique des processus impliqués dans la perception des expressions vocales émotionnelles.

La *fonctionnalité* de la perception est une dimension centrale du modèle en lentille de Brunswik. Pour que la perception – ou la communication, dans le cas des expressions vocales émotionnelles – soit fonctionnelle, il est nécessaire que l'utilisation des indicateurs disponibles (les caractéristiques vocales) par un observateur (auditeur) soit conforme à la relation objective entre ces indicateurs et l'objet de l'environnement qui est perçu/identifié (l'émotion exprimée). Dans la perspective de Brunswik, les relations entre les indicateurs (*cues*) disponibles et les objets perçus sont *probabilistes*. Ces relations probabilistes sont apprises par les individus au cours de leur développement. La présence d'un indicateur spécifique correspond donc pour un individu à une certaine probabilité qu'un objet donné soit présent dans l'environnement.

De plus, un objet ou une dimension de l'environnement correspond habituellement à une *multitude d'indicateurs* différents qui sont liés plus ou moins systématiquement/probablement à cet objet/dimension. La multiplication des indicateurs associés (plus ou moins systématiquement) à des objets perçus implique, d'une part, que certains indicateurs peuvent être présents dans certains cas et absents dans d'autres cas et qu'une même perception peut être réalisée sur la base d'indicateurs partiels. A l'extrême on peut donc imaginer que deux séries totalement différentes d'indicateurs puissent donner lieu à une même perception. D'autre part, de la multiplication des indicateurs résulte aussi la présence d'indicateurs redondants. La *redondance* des indicateurs assure la robustesse de la perception, malgré la relation non-déterministe entre les objets et leurs indicateurs.

L'application de cette perspective à la communication vocale des émotions conduit à considérer certaines possibilités qui ne sont pas nécessairement en accord avec les "modèles" (généralement implicites) qui ont cours dans ce domaine d'étude. La relation non-déterministe entre les émotions exprimées et les caractéristiques vocales contraste en particulier avec la démarche qui vise à définir un profil acoustique spécifique pour chaque type d'émotion exprimée. Dans une perspective brunswikienne, l'association entre un ensemble de caractéristiques vocales et une certaine émotion s'exprimerait en terme de probabilité et non de manière déterministe. Dans ce cadre, plusieurs

profils acoustiques différents pourraient notamment correspondre à une même émotion exprimée et pourraient également communiquer cette émotion. De plus, dans cette perspective, les caractéristiques vocales contribuent *indépendamment* à la perception des émotions exprimée en fonction de leur probabilité d'association avec ces émotions.

La *représentativité* des designs expérimentaux est un autre concept central dans la perspective proposée par Brunswik. Une condition expérimentale est dite *représentative* lorsqu'elle inclut des configurations d'indicateurs qui surviennent également hors du cadre expérimental. Selon Brunswik, les résultats obtenus dans des conditions non-représentatives ne peuvent être généralisés en dehors du cadre expérimental. Avec cette notion, Brunswik s'attaque à la démarche classique de la psychologie expérimentale qui consiste à manipuler indépendamment plusieurs variables dans un design factoriel complet; il est possible dans ce cas qu'une partie des conditions qui résultent de cette manipulation ne soient pas représentatives, au sens où elles ne correspondraient pas à des conditions qui existeraient également dans la réalité quotidienne des individus. De telles conditions peuvent notamment "mettre en échec" des processus perceptifs qui sont néanmoins fonctionnels dans des situations représentatives.

Cette notion peut s'appliquer à plusieurs aspects de l'étude de la communication vocale des émotions. Le paradigme du contenu verbal constant a par exemple été critiqué sur une base similaire. Dans ce paradigme, le contenu verbal utilisé pour exprimer différentes émotions est en principe "neutre". Le postulat de neutralité du contenu verbal est toutefois problématique; il est en effet possible qu'une phrase telle que "ils ont acheté une nouvelle voiture" ou "sa fiancée est venue en avion" (exemples tirés de Mozziconacci, 1998) puissent être, pour un individu donné, plus proche d'un certain type d'émotion exprimée que d'un autre. Il est donc possible que le paradigme du contenu verbal constant puisse générer des configurations où le contenu verbal et l'émotion exprimée sur le plan non-verbal seraient potentiellement en contradiction. Ces expressions ne seraient alors pas représentatives d'un cadre de communication usuel, mais correspondrait plutôt à une situation assez particulière où des informations ambiguës seraient communiquées.

A un autre niveau, la notion de représentativité peut être appliquée aux caractéristiques vocales qui sont en partie redondantes dans les expressions vocales émotionnelles produites par des locuteurs ou des acteurs. Dans le cas des expressions vocales, la redondance partielle de différentes caractéristiques vocales – tels que la hauteur ou l'intensité – s'explique par la dépendance de ces caractéristiques relativement à un mécanisme de production commun. Lorsque, par exemple, l'effort vocal est augmenté, la hauteur moyenne et l'intensité moyenne, ainsi que la proportion d'énergie dans les hautes fréquences augmentent. Un autre exemple de cette sorte est présenté par Juslin &

Laukka (2003) qui évoquent la colinéarité de variables reflétant des aspects de la qualité vocale. Ils notent qu'une voix "tendue" – reflétant une tension et une constriction du pharynx et un rétrécissement du tractus vocal – produit à la fois une élévation du premier formant, une bande de fréquence plus étroite pour ce formant et une élévation de l'énergie dans les hautes fréquences; alors qu'une voix "détendue" – correspondant à un pharynx détendu et à une pression subglottale plus faible – produit simultanément un abaissement du premier formant et un élargissement de sa bande de fréquence, une articulation moins précise et une diminution de la proportion d'énergie dans les hautes fréquences. Cette observation est particulièrement importante dans le contexte des études qui tentent de manipuler indépendamment différents aspects des expressions vocales émotionnelles. Relativement à des expressions "naturelles", certaines configurations de caractéristiques vocales manipulées ne sont pas représentatives, au sens où elles correspondent à des configurations qui ne pourraient être produites par un locuteur ou un acteur.

Les paragraphes qui précèdent font référence à la *perception*. Le modèle en lentille et les concepts qui lui sont associés ont été effectivement développés par Brunswik dans le cadre de l'étude des phénomènes perceptifs (en particulier dans le cadre de l'analyse des constances perceptives dans le domaine visuel). Toutefois, les phénomènes qui ont été examinés par Brunswik comprennent à la fois des perceptions élémentaires et la formation de représentations abstraites relativement complexes. La distinction entre "perception" et "pensée" est évidemment très ancienne. Brunswik (1956) avait entrepris de formaliser cette distinction en opposant l'*intuition* à l'*analyse*. On trouve toutefois aussi chez Brunswik l'idée, encore peu élaborée, d'un continuum entre ces deux types de cognitions. Hammond (2001) a repris et développé cette notion de *continuum cognitif*. Les deux extrêmes de ce continuum correspondent à l'*intuition* – qui comprend un contrôle cognitif faible, un traitement rapide de l'information, un niveau de conscience faible – et à l'*analyse* – qui comprend un contrôle cognitif élevé, un traitement de l'information lent et un niveau de conscience élevé. Curieusement, le modèle en lentille et les concepts développés par Brunswik ont été appliqués plus souvent à des problématiques relevant de l'analyse cognitive (e.g. analyses de jugements ou processus de prise de décision) qu'à l'étude de phénomènes perceptifs plus élémentaires. La section ci-dessous présente un très bref survol de quelques applications du modèle en lentille qui ont été effectuées par différents auteurs.

1.3.3 Applications du modèle en lentille de Brunswik à la recherche en psychologie

La proposition de Scherer (1978, 1982, 2003) concernant l'utilisation du modèle en lentille pour la recherche sur la communication vocale des émotions n'a pas rencontré un grand succès. A notre connaissance un seul auteur a repris cette proposition dans le domaine de la communication non-

verbale des émotions. Il s'agit de Juslin (1998) qui a utilisé le paradigme du modèle en lentille pour analyser la communication des émotions par la musique. En revanche, le modèle en lentille a été exploité dans d'autres domaines de recherche; en particulier pour *l'analyse de jugements* ('judgment analysis'), mais également dans le domaine de la *perception interpersonnelle* ('interpersonal perception').

Le modèle en lentille de Brunswik a été employé et soutenu relativement rapidement par d'autres auteurs, en particulier par Hammond (1955) qui est encore aujourd'hui l'un des plus fervents défenseurs de ce paradigme. La première application du modèle en lentille par Hammond (1955) a été, déjà, effectuée dans le domaine de l'analyse de jugements ('judgment analysis' ou 'decision making'). Hammond a utilisé ce paradigme pour étudier (décomposer) les jugements cliniques (diagnostiques) de psychiatres et de psychologues. L'approche brunswikienne a par la suite été régulièrement appliquée à l'étude des jugements diagnostiques. La raison probable de l'intérêt pour cette approche dans ce domaine est donnée par Wigton (Wigton, 2001, p. 379): *"Because it is easy to find medical judgment tasks with multiple fallible indicators, they have been excellent subjects of lens model research."* (v. aussi Wigton, 1996). A la suite de Hammond, d'autres auteurs ont appliqué le paradigme de Brunswik à d'autres types de jugements dans lesquels des indicateurs multiples et "faillibles" (probabilistiques) interviennent. L'exemple le plus typiquement cité dans ce domaine est probablement l'analyse des jugements concernant les prévisions météorologiques (dans le domaine plus large de l'étude des "forecasting skills", v. Stewart, Moninger, Heideman, & Reagan-Cirincione, 1992, ou Stewart & Lusk, 1994).

Holzworth (2001) a présenté une brève revue de la recherche, relative aux analyses de jugements, effectuée dans une perspective brunswikienne. Dans cette revue, cet auteur relève que l'intérêt des études dévolues à l'analyse de jugements c'est relativement rapidement orienté vers les problématiques de la perception interpersonnelle (v. Hammond, Wilkins, & Todd, 1966). Dans ce domaine, Albright & Malloy (2001) vont jusqu'à affirmer: *"All research on interpersonal perception is Brunswikian, because the use of real people as targets of judgment invokes the principle of representative covariation. Because Brunswik was the first to conduct a theoretically based and comprehensive […] study of social perception, he was the originator of the interpersonal approach to social perception research."* (Albright & Malloy, 2001, pp. 330-331). Sans nécessairement reprendre cette affirmation à notre compte, nous présentons brièvement ci-dessous trois exemples de recherches dans le domaine de la perception interpersonnelle qui font explicitement référence au modèle en lentille.

Dans une étude publiée en 1994, Gifford a présenté une analyse de la communication non-verbale d'un ensemble de 'dispositions interpersonnelles' correspondant aux traits suivants: dominant - soumis, agréable - désagréable, introverti-extraverti et arrogant-ingénu. L'auto-évaluation de ces traits a été obtenue pour 60 participants par le biais d'un questionnaire. Un groupe de juges indépendant a effectué l'évaluation des même traits pour ces participants sur la base d'enregistrements vidéo (muets) filmés dans des situations de conversation libre. Différents aspects relatifs aux mouvements, à la posture et aux expressions faciales observables dans ces enregistrements vidéo ont été codés par un groupe d'experts. Dans cette étude, Gifford a pu examiner, d'une part, la relation entre les traits auto-évalués et les caractéristiques non-verbales codées (encodage) et, d'autre part, la relation entre les traits attribués par le groupe de juges et les caractéristiques non-verbales codées (décodage); ainsi que la relation entre les traits auto-évalués et les traits attribués par les observateurs. Cette procédure lui a notamment permis de mettre en évidence certains aspects non-verbaux susceptibles de communiquer (encoder et décoder) l'introversion et l'extraversion.

En 2001, Reynolds & Gifford ont publié une étude dans laquelle ils ont examiné la perception de l'intelligence sur la base d'enregistrement audio-vidéo, d'enregistrements vidéo (sans audio) et d'enregistrements audio (sans vidéo) d'un groupe de 30 participants. Dans cette étude, l'intelligence des participants a été évaluée à l'aide d'un test (WPT, un test dérivé du WAIS-R), différents groupes d'observateurs ont ensuite donné une évaluation de l'intelligence des participants sur la base des enregistrements audio, vidéo ou audio-vidéo des participants qui répondaient à des questions de l'expérimentateur. Des groupes indépendants de juges ont évalué plusieurs aspects non-verbaux (par exemple les sourires, la taille corporelle, l'attractivité) et des aspects para-verbaux (par exemple le débit de parole, l'intensité vocale, les hésitations). Dans cette étude, la correspondance entre l'intelligence mesurée par le test et l'intelligence attribuée par les observateurs était faible, bien que différents indicateurs – en particulier – des indicateurs para-verbaux – semblent avoir été utilisés de manière relativement cohérente par les observateurs au décodage et bien que certains indicateurs soient parvenus à rendre compte d'une partie de la variance de l'intelligence mesurée. Cette étude a donc notamment permis de mettre en évidence que certains aspects non-verbaux corrélés à l'intelligence (mesurée par un test) ne sont pas exploités par les observateurs pour inférer l'intelligence.

Un autre exemple de l'utilisation explicite du paradigme de Brunswik peut être trouvé dans une étude de Bernieri & Gillis (2001). Ces auteurs ont évalué la perception de la 'qualité du rapport' entre deux participants dans le cadre d'une interaction libre. Le critère pour l'évaluation du rapport entre les deux participants a été établi sur la base des auto-évaluations fournies par chaque

participant relativement au rapport perçu avec son partenaire. Un groupe de juges a été recruté pour évaluer le rapport (plus ou moins bon/mauvais) entre les participants formant une paire. Par ailleurs, des experts ont codés différents aspects comportementaux (e.g. les contacts oculaires, la proximité physique, des aspects liés à la posture, les sourires) pour toutes les paires de participants. Cette procédure a permis à Bernieri & Gillis d'évaluer les comportements qui sont indicatifs de la qualité du rapport vécu par les paires de participants et, d'autre part, d'évaluer la correspondance de ces indicateurs de la qualité du rapport vécu avec les indicateurs qui permettent de prédire les jugements relatifs à la qualité du rapport effectués par des observateurs.

Dans les deux premiers exemples présentés ci-dessus, le modèle en lentille est appliqué à deux problématiques assez proches, sur le plan de leur forme, de la problématique de la communication vocale des émotions. Dans les deux cas, des attributs abstraits (hypothétiques) des participants – des traits de personnalité ou l'intelligence – sont évalués sur la base d'un critère – l'auto-évaluation ou un test – défini par les auteurs. Des jugements relatifs aux attributs des participants sont obtenus en interrogeant un groupe de juges; et un ensemble de caractéristiques comportementales sont codées dans le but d'évaluer, d'une part, leurs relations avec le critère défini pour l'attribut étudié et, d'autre part, leurs relations avec les jugements obtenus. La relation entre les caractéristiques comportementales est les jugements peut alors être comparée avec la relation entre ces caractéristiques et le critère défini pour l'attribut étudié. La problématique examinée dans le dernier exemple est légèrement plus complexe, au sens où l'objet de la communication n'est pas l'attribut d'une seule personne, mais l'attribut d'une relation entre deux personnes.

Ces exemples démontrent que le paradigme du modèle en lentille a été récemment exploité pour l'étude de plusieurs problématiques voisines de la question de la communication vocale des émotions. Un objectif de la recherche présenté dans cette thèse est de mettre en avant l'utilisation de ce paradigme également dans le domaine de la recherche sur la communication vocale émotionnelle où, malgré les propositions de Scherer, ce paradigme n'a pas été exploité à ce jour.

1.4 Rôle de l'intonation dans la communication émotionnelle

Lorsque nous parlons de communication *vocale* ou *non-verbale*, nous faisons référence à l'ensemble des caractéristiques de la voix et de la parole qui ne relèvent pas des domaines phonétique, sémantique et syntaxique, c'est-à-dire à tous les aspects des expressions orales qui ne seraient pas représentés dans une transcription écrite. Ceci recouvre un grand nombre de caractéristiques qui sont génériquement désignées par différents termes tels que 'qualité vocale', 'timbre vocal', 'intonation' ou 'prosodie'. Ces termes génériques ne sont malheureusement pas unanimement et précisément définis. Le terme 'prosodie' peut par exemple recouvrir la totalité des caractéristiques

non-verbales dans une certaine acception, mais peut aussi désigner uniquement les phénomènes relatifs à l'intonation – à l'exclusion des aspects liés à la qualité vocale – dans d'autres définitions.

La distinction entre 'intonation' et 'qualité vocale' est toutefois particulièrement saillante dans le domaine de l'étude des phénomènes vocaux. Le terme 'intonation' recouvre des aspects indiscutablement suprasegmentaux – mélodiques et rythmiques – des expressions vocales qui ont été examinés surtout dans une tradition de recherche linguistique. Le terme 'qualité vocale' recouvre quant à lui des aspects liés à différents modes de phonation et à la forme du tractus vocal qui conditionne les résonances. Ces aspects peuvent se modifier, évoluer, au cours d'une expression; dans ce sens ils relèvent également du domaine suprasegmental. Pourtant, contrairement aux phénomènes rythmiques et mélodiques, ils peuvent être identifiés déjà au niveau segmental. La qualité vocale est essentiellement examinée dans un courant de recherche qui s'intéresse aux pathologies vocales. Sur le plan scientifique, la qualité vocale et l'intonation relèvent donc de spécialisations différentes et sont rarement considérées simultanément.

Dans cette section, nous allons aborder la question du rôle de l'*intonation* dans la communication vocale des émotions. Quelques éléments de définition et un survol des modèles qui ont été proposé pour la description et l'analyse de l'intonation seront d'abord présentés. Un certain nombre de propositions et de données relatives à l'influence des émotions sur l'intonation seront ensuite examinées.

1.4.1 Définition de l'intonation dans la littérature

Dans les travaux des psychologues, l'intonation est le plus souvent définie sur le plan opérationnel; on ne trouve pratiquement pas de modèles ou de définitions très élaborées. Dans le domaine de la linguistique en revanche, l'intonation est un objet d'étude qui a retenu l'attention de nombreux auteurs et a donné lieu à différents modèles et définitions. Dans une définition publiée en 1970, Léon & Martin évoquent les usages multiples du terme "intonation". Ils écrivent: *"La plupart des auteurs ne définissent pas l'intonation. Lorsqu'il s'agit d'en préciser la nature, certains emploient le terme tantôt d'une façon étroite (les variations de hauteur uniquement), tantôt d'une façon large (incluant dans l'acception du terme les paramètres d'intensité et de durée). Tant du point de vue de la production (physique ou physiologique) que de celui de la réception (perceptive), les 3 paramètres: durée, intensité et hauteur sont étroitement liés; même si l'on étudie que les fluctuations de la ligne mélodique, on est obligé de tenir compte des autres variables qui l'accompagnent. Cependant statistiquement, les variations de hauteur apparaissent comme les plus importantes pour la perception de l'intonation et ce sont d'elles que nous traiterons surtout ici."* (Léon & Martin, 1970, p. xv). On retrouve pratiquement la même définition chez Cruttenden

(Cruttenden, 1986, pp. 2-3): *"The prosody of connected speech may be analysed and described in terms of the variation of a large number of prosodic features. There are, however, three features which are most consistently used for linguistic purposes either singly or jointly. These three features are pitch, length, and loudness. […] Pitch is the prosodic feature most centrally involved in intonation and it is with this feature that I shall be principally concerned in this book."*

L'intonation est donc définie comme recouvrant les variations de durée, de hauteur et d'intensité. Ces trois paramètres ne sont pas indépendants, mais la plupart des auteurs considèrent la hauteur perçue comme l'élément "le plus important" ou "le plus central" et restreignent la définition de l'intonation aux variations de la hauteur perçue des expressions. Cet usage restreint du terme intonation est très fréquent dans les études linguistiques de l'intonation; la majorité d'entre elles s'attachent en conséquence essentiellement à décrire les fluctuations de la hauteur perçue de la parole.

1.4.2 Transcriptions et modèles de l'intonation

Des tentatives de transcription et de formalisation de l'intonation apparaissent déjà au 18ème et 19ème siècle; elles sont alors souvent effectuées dans un objectif pédagogique (pour une revue des premiers travaux effectués dans ce domaine v. Crystal, 1969, pp. 20-40). Au cours du 20ème siècle de nombreux systèmes de *transcription de la hauteur perçue* sont élaborés dans le cadre d'approches linguistiques de l'étude de l'intonation. En 1970, Léon & Martin distinguent notamment six types de transcriptions: les "transcriptions directes", les "transcriptions musicales", les "transcriptions de patrons intonatifs", les "transcriptions figurant l'intonation et l'accentuation", les "transcriptions figurant des niveaux d'intonation" et les "transcriptions par courbes et niveaux" (Léon & Martin, 1970, pp. 26-32).

Chaque système de transcription repose sur différents choix relatifs essentiellement à la finesse des changements décrits et au type de description (en tons ou en courbes). La finesse des changements de hauteur pris en compte dépend du nombre de niveaux considérés par un système de transcription. Le nombre de niveaux varie de 2 (haut et bas) à l'ensemble des niveaux descriptibles sur une portée musicale. La notation peut être réalisée en tons, dans ce cas un niveau de hauteur est indiqué pour chaque syllabe, avec des mouvements implicites d'une syllabe à l'autre. Alternativement, la notation peut être effectuée en courbes, dans ce cas un mouvement est indiqué pour certaines syllabes, lorsqu'un changement de hauteur est perceptible. Ces choix conditionnent évidemment la forme finale de la transcription. Les tons ou les courbes sont conçus comme les éléments d'une 'grammaire intonative' et le but des transcriptions est souvent de répertorier pour une langue donnée les combinaisons possibles de ces éléments et leurs significations.

Depuis quelques années, le système ToBI (Tones and Break Indices system, Silverman et al., 1992) s'est profilé comme le système de référence pour la transcription de l'intonation de la langue anglo-américaine. Le système ToBI a été développé dans le but d'unifier les méthodes utilisées par les chercheurs de différents domaines pour décrire l'intonation (la hauteur perçue). Ce système propose de décrire l'intonation par une succession de tons (hauts/bas) placés sur des cibles spécifiques dans les expressions verbales; un ton est placé notamment sur chaque syllabe accentuée (pitch accents) et à chaque frontière intonative (phrasal tones, final boundary tones). Ce système a été assez largement adopté pour la transcription de l'intonation de l'anglais et a été adapté à d'autres langues (e.g. pour la description du Coréen: K-ToBI, pour l'allemand G-ToBI). Mertens a développé un système similaire pour la description de l'intonation du français (Mertens, 1987).

Le système ToBI n'est pas athéorique. Il est basé sur un modèle de séquence tonale (Tone Sequence Model) qui a été originalement proposé par Pierrehumbert (1980) pour l'anglo-américain. Le modèle de la génération de l'intonation de Pierrehumbert (1980) vise à décrire des règles intonatives fondamentales, ainsi qu'un ensemble de règles de transformation qui seraient capables de générer les séquences de tons réalisées dans différents contextes (pour une synthèse de ce modèle v. Cruttenden, 1986, pp. 67-72).

Les modèles de séquence tonale ont fait l'objet de critiques relativement à certaines de leurs implications théoriques et à certaines de leur limites concernant la description des contours de hauteur. Ces critiques ont été formulées surtout par les défenseurs d'une autre perspective qui conçoivent le contour intonatif comme le résultat de la superposition de plusieurs composantes. Le modèle type de cette approche est le modèle à deux composantes (e.g. Fujisaki, 1988, 1997; Ladd, 1983a, 1983b), où la première composante correspond à un mouvement de la hauteur à long terme – par exemple une "ligne de déclinaison" (un niveau de base qui décline lentement du début à la fin d'une production vocale) – et où la deuxième composante correspond aux variations de hauteur à plus court terme, tels que les accents locaux qui se superposent à la ligne de déclinaison.

Dans les modèles à plusieurs composantes (Superpositional Models), des mouvements (contours) de hauteur constituent les éléments de base de l'intonation, alors que les modèles de séquence tonale conçoivent l'intonation comme une séquence de tons placés sur des cibles. Plus ou moins explicitement, ces deux types de modèles défendent donc des hypothèses différentes relativement à la perception de l'intonation qui, selon les modèles à plusieurs composante, serait dépendante de la perception de mouvements (superposés) de la F0 et, selon les modèles de séquence tonale, dépendrait essentiellement de la perception de la hauteur relative de certains points cibles dans le contour. Les défenseurs des modèles à plusieurs composantes ont plus particulièrement mis en

cause deux limites des modèles de séquence tonale: premièrement, le fait que ces modèles ne puissent rendre compte efficacement des effets d'anticipation ou de reprise (e.g. Kutik, Cooper, & Boyce, 1983) qui sont observables dans la production des contours d'intonation et, deuxièmement, la nécessité pour ces modèles d'introduire un concept ad-hoc – la notion de "downstep" – permettant de rendre compte de la déclinaison progressive de la hauteur à long terme (Ladd, 1983b).

Les modèles à plusieurs composantes comportent également leurs propres limites. Les contours superposés qui composent l'intonation dans cette perspective sont des abstractions qui ne se prêtent pas aisément à l'association avec des catégories ou des significations linguistiques. De plus, les différentes composantes ne peuvent être définies que relativement les unes aux autres. Dans un modèle à deux composantes par exemple, l'une des composante (en principe la ligne de déclinaison) doit être fixée en fonction de critères prédéfinis, de manière à définir l'autre composante (dans ce cas les variations locales).

Les modèles de l'intonation évoqués ci-dessus ont été développés de manière à rendre compte de l'intonation linguistique. La finalité de ces modèles est, en principe, de rendre compte de la génération de l'intonation pour une langue spécifique, c'est-à-dire de formuler une "grammaire de l'intonation" pour cette langue (v. par exemple Pierrehumbert, 1980, pour l'anglais; ou t'Hart, Collier, & Cohen, 1990, pour le hollandais). La relation de l'intonation avec d'autres catégories linguistiques – syntaxiques, sémantiques ou pragmatiques – est avant tout examinée dans ce cadre; les aspects dits "expressifs" de l'intonation (non-linguistiques ou para-linguistiques) sont en général considérés comme secondaires. En conséquence, il n'est pas établi que ces modèles de l'intonation puissent rendre compte également des variations intonatives liées à des aspects non-linguistiques tels que la communication émotionnelle.

La communication par l'intonation d'aspects qui n'entrent pas directement dans le domaine de l'étude de la langue a pourtant retenu l'attention d'un grand nombre d'auteurs. La plupart des ouvrages consacrés à l'intonation (v. par exemple Halliday, 1970; Lacheret-Dujour & Beaugendre, 1999; O'Connor & Arnold, 1973) incluent un chapitre ou une section consacrés à la "fonction expressive" de l'intonation; c'est-à-dire à l'utilisation de l'intonation pour la communication des émotions et des attitudes. Indépendamment des propositions formulées dans ces ouvrages, "l'intonation émotionnelle" a été étudiée dans différents contextes de recherche. Des études de neuropsychologie clinique ont par exemple été consacrées à différencier les processus de traitement associés à l'intonation linguistique d'une part et à l'intonation "émotionnelle" d'autre part (e.g. Heilman, Bowers, Speedie, & Coslett, 1984; Pell, 1998; Ross, 1981; van Lancker & Sidtis, 1992). Dans un autre domaine, des études qui se sont intéressée à la production ou à la perception

prélinguistique de l'intonation chez les jeunes enfants suggèrent que différentes significations émotionnelles peuvent être communiquées avant l'acquisition du langage (Fernald, 1991, 1992, 1993; Papousek, Papousek, & Symmes, 1991). L'intonation linguistique et l'intonation émotionnelle sont dans toutes ces approches considérées comme relevant de processus indépendants. L'existence éventuelle d'interactions au niveau de la production et/ou de la perception de ces deux types d'intonation est, à notre connaissance, beaucoup plus rarement considérée.

Dans la section qui suit, différentes propositions, ainsi que quelques résultats empiriques relativement à l'influence de l'intonation sur la perception et l'expression des émotions sont présentés. Deux travaux qui ont abordé plus spécifiquement la question de l'indépendance/interaction entre l'intonation linguistique et l'intonation émotionnelle (Scherer, Ladd, & Silverman, 1984; Uldall, 1964) sont notamment décrits.

1.4.3 Effets observés et effets postulés de l'émotion sur l'intonation

Une tradition de recherche déjà évoquée ci-dessus rassemble des auteurs qui se sont efforcés d'identifier et de décrire des contours d'intonation spécifiques qui correspondraient à des émotions ou à des attitudes données (v. notamment Fonagy & Magdics, 1963; Halliday, 1970; O'Connor & Arnold, 1973). Les descriptions de Fonagy & Magdics (1963) illustrent bien ce type d'approche. Ces auteurs décrivent l'évolution de la hauteur perçue par une succession de tons sur une portée musicale pour différents énoncés qui correspondent à différentes situations émotionnelles. L'énoncé *"Comme je suis heureuse de te voir! Je ne pensais pas te rencontrer!"* est par exemple utilisé pour illustrer un contour typique de joie. Frick (1985) a émis une série de critiques relativement à ce type d'approche. Selon cet auteur, l'utilisation d'un exemple contextualisé dans le but de démontrer qu'un contour spécifique possède une signification émotionnelle spécifique est problématique car, premièrement, le contenu verbal contient souvent la signification que le contour devrait en principe transmettre et, deuxièmement, le lecteur peut ajouter à la description du contour fournie par l'auteur d'autres éléments prosodiques (non-spécifiés par l'auteur) qui contribueront à produire l'impression émotionnelle.

Une étude récente de l'intonation émotionnelle Mozziconacci (1998) s'est inscrite dans cette approche en tentant d'identifier empiriquement des contours de hauteur spécifiques pour plusieurs états émotionnels. Mozziconacci (1998) a défini (a priori) un certain nombre de contours (configurations d'accents), basés sur le modèle de l'intonation développé par l'IPO (d'après les propositions de t'Hart et al., 1990), puis a tenté d'associer ces contours avec des expressions émotionnelles. Mozziconacci (1998) n'est pas parvenue à identifier des contours (configuration d'accents) qui seraient systématiquement associés avec différentes émotions exprimées, mais elle

défend néanmoins l'idée que des caractéristiques "optimales" liées à l'intonation – des contours spécifiques associés à des variations du niveau de hauteur, de l'étendue de hauteur et de la durée – permettraient de communiquer des émotions spécifiques.

A notre connaissance, très peu d'auteurs ont remis en cause explicitement cette conception du rôle de l'intonation dans la communication émotionnelle. On trouve toutefois certaines exceptions, Pakosz (1983) a, par exemple, défendu une théorie qui postule que seul le niveau d'activation est transmis (exprimé et perçu) sur le plan de l'intonation des expressions vocales. Selon cet auteur, d'autres indices contextuels (ou des expressions faciales) doivent être mis en jeu pour communiquer des émotions spécifiques.

La plupart des études "classiques" qui ont tenté d'examiner le rôle de l'intonation dans le cadre de la reconnaissance des expressions émotionnelles ont cependant mis en évidence une participation de différentes caractéristiques vocales liées à l'intonation dans la communication des émotions. Ces études ont pour la plupart été réalisées non pas dans l'objectif de définir des contours spécifiques pour des émotions spécifiques, mais plutôt afin de tester l'importance relative de l'intonation et de la qualité vocale dans la communication émotionnelle ou afin d'évaluer des associations statistiques entre certaines caractéristiques liées à l'intonation et différentes émotions perçues. Dans les paragraphes qui suivent quelques illustrations de ce type d'approche sont présentées.

Dès les années soixante, différents auteurs (e.g. Scherer et al., 1985) ont tenté de séparer la contribution respective de la qualité vocale et de l'intonation à la communication vocale des émotions. Ils ont utilisé différentes méthodes de "dégradation" du signal acoustique destinées à affecter sélectivement la qualité vocale (par des techniques de filtrage du signal acoustique) ou l'intonation (en altérant la séquence temporelle du signal). Dans d'autres travaux, des techniques de re-synthèse vocale ont été utilisées afin de modifier sélectivement différents aspects de l'intonation.

Une étude classique dans ce domaine a été publiée par Lieberman & Michaels (1962). Ces auteurs ont utilisé des expressions correspondant à 8 "modes émotionnels"[2] produits pour plusieurs énoncés anglais. Dans cette étude, 85% des "modes émotionnels" ont été reconnus correctement par un groupe d'auditeurs lorsque les enregistrements originaux leur étaient présentés. Les auteurs ont re-synthétisé la fréquence fondamentale de ces expressions sur une voyelle fixe, cette opération conserve l'information relative à l'évolution dans le temps de la fréquence fondamentale tout en supprimant les modifications spectrales (i.e la qualité vocale) et les variations d'amplitude des

[2] Il s'agit en fait des catégories suivantes: 1. 'bored statement', 2. 'confidential communication', 3. 'question expressing disbelief', 4. 'message expressing fear', 5. 'message expressing happiness', 6. 'objective question', 7. 'objective statement', 8. 'pompous statement'

expressions originales. Pour ces expressions produites par synthèse avec les contours de F0 originaux, la reconnaissance correcte globale des différents "modes émotionnels" était réduite à 44%. L'ajout du contour d'amplitude au contour de F0 n'a permis d'augmenter que de 3% (à 47%) la proportion de réponses correctes. En revanche, le "lissage" ("smoothing", i.e. la suppression des variations à court terme) des contours de F0, dans la même étude, a fait chuter le pourcentage de reconnaissance correcte : 38% des expressions de synthèse avec un contour lissé à 40 ms et, seulement, 25% des expressions de synthèse avec un contour lissé à 100 ms étaient encore reconnues correctement. Cette étude indique que le contour de F0 et les informations spectrales contribuent à la reconnaissance des caractéristiques non-verbales exprimées. Elle souligne également que les variations à court terme du contour de fréquence fondamentale jouent un rôle important dans ce domaine. Lieberman & Michaels (1962) attribuent la diminution de la reconnaissance introduite par le lissage des contours à la présence de microperturbations ("jitter") dans certaines expressions. Selon eux, ces microperturbations permettraient notamment de différencier certaines expressions de peur et de joie qui seraient confondues uniquement lorsque les contours de F0 sont lissés. Dans cette étude, les variations à plus long terme des contours de F0 (i.e. les variations qui sont habituellement prise en compte dans la description de l'intonation) contribuent également à la reconnaissance des "modes émotionnelles", mais ne permettent de différencier correctement que 25% des "modes émotionnelles".

Dans une série de trois études, Ladd, Silverman, Tolkmitt, Bergmann, & Scherer (1985) ont évalué l'effet d'une manipulation du contour de F0 ("uptrend" versus "downtrend") et de l'étendue de F0 (graduellement augmentée) synthétisées sur des expressions produites par différents locuteurs ; pour l'un de ces locuteurs deux types de qualité vocale ont été enregistrées ("normale" et "rauque"). Leurs résultats indiquent que le type de contour, l'étendue de la F0 et la qualité vocale affectent les jugements émotionnels de manière indépendante. Ils observent toutefois un effet plus important de l'étendue de la F0 – dont l'augmentation progressive influence graduellement l'intensité des attributions émotionnelles – relativement à la forme des contours de F0 qui affectent les attributions émotionnelles plus faiblement et moins systématiquement dans cette étude.

Uldall (1964) a publié une autre étude pionnière dans laquelle 16 contours de F0 stylisés ont été appliqués systématiquement sur 5 phrases prononcées par un locuteur. Uldall décrit, pour chaque phrase, et pour chaque contour les tendances qui émergent statistiquement des jugements fournis par un groupe d'auditeurs à l'aide d'une série d'échelles sémantiques différentielles (semantic differential scales). Ses résultats démontrent que la signification qui se dégage de différents types de contours varie en fonction de la phrase. Elle trouve par exemple que le contour qui correspond à une déclinaison avec une pente faible et un niveau bas se classe comme 'déplaisant', 'autoritaire' et

exprimant une émotion 'faible' lorsqu'il est appliqué aux 2 phrases interrogatives et à la phrase déclarative. Alors que le même contour se classe comme 'déplaisant', 'autoritaire', et exprimant une émotion 'forte' lorsqu'il est appliqué à la phrase formulée à l'impératif. Par ailleurs, Uldall identifie une série de propriétés des contours qui sont liées aux 3 dimensions qu'elle a dégagées à partir des jugements des auditeurs.

Scherer et al. (1984) ont effectué une étude dans laquelle ils ont également mis en évidence une interaction entre une caractéristique du contour de F0 et une catégorie linguistique. Dans cette étude, les attributions émotionnelles ont été différemment influencées par la forme finale du contour de F0 (montée versus descente) dans des phrases interrogatives en fonction de leur type grammatical. Les questions qui appellent une réponse oui/non ("yes/no questions") sont jugées agressives ou provocantes lorsqu'elles comportent un contour final descendant, alors que les questions qui débutent par un mot interrogatif (qui, quoi, quand, etc., "WH questions") sont jugées neutres lorsque leur contour final est descendant. Dans cette étude, les auteurs ont explicitement testé 2 hypothèses (ou modèles) concernant la manière dont différentes caractéristiques vocales influencent l'attribution émotionnelle. Ils ont montré que d'une part certaines catégories telles que le contour final – montant versus descendant – influencent les attributions émotionnelles en fonction de l'interaction avec d'autres catégories, notamment linguistiques. D'autre part, ils ont pu montrer que la variation continue de certaines caractéristiques affecte directement les attributions émotionnelles. L'augmentation (continue) de la fréquence fondamentale moyenne corrèle par exemple avec les jugements du degré d'activation émotionnelle des locuteurs. Ces auteurs ont proposé que les variables qui affectent les attributions émotionnelles de manière continue reflèteraient surtout l'activation physiologique liée à la réaction émotionnelle du locuteur. Alors que les variables catégorielles qui affectent l'intonation en interaction avec des catégories linguistiques tendraient à signaler plutôt des attitudes du locuteur (par exemple amicale ou réprobatrice).

Les exemples présentés ci-dessus illustrent le rôle attribué à l'intonation dans une partie de la littérature sur la communication vocale des émotions. Les 2 premières études présentées (Ladd et al., 1985; Lieberman & Michaels, 1962) indiquent essentiellement que des aspects isolés de l'intonation parviennent à transmettre une partie de l'information émotionnelle présente dans les expressions originales. Les deux autres études (Scherer et al., 1984; Uldall, 1964) cherchent à définir plus précisément de quelle manière certaines propriétés de l'intonation – dans les deux cas des aspects des contours de F0 – affectent les attributions émotionnelles. Dans ce type d'étude, les résultats indiquent souvent que différentes caractéristiques de l'intonation (plus spécifiquement des contours de F0) reçoivent des attributions émotionnelles différentes en interaction avec le contenu linguistique (syntaxique ou sémantique) des énoncés.

1.5 Problématiques et questions de recherche

Plusieurs problématiques émergent de la revue de la littérature présentée ci-dessus. Les principales questions que nous avons retenues sont présentées dans ce qui suit.

1. La scission entre études d'encodage et études de décodage préjudicie considérablement le développement des connaissances relatives à la *communication* vocale des émotions. Dans le domaine de l'encodage, les expressions étudiées sont souvent sélectionnées sur la base de la reconnaissance par des auditeurs de l'émotion exprimée, mais les caractéristiques acoustiques des expressions ne sont généralement mises en relation qu'avec les émotions exprimées et non avec les émotions perçues. Les chercheurs qui se sont intéressés à la reconnaissance des émotions ont quant à eux mis en évidence la capacité des auditeurs à reconnaître différentes émotions, en particulier lorsqu'elles sont simulées par des acteurs, mais ils n'ont que très rarement tenté de décrire les propriétés des expressions identifiées par les auditeurs comme correspondant à différentes émotions. De ce fait, il n'existe que peu de données relativement aux caractéristiques vocales qui interviennent dans le décodage.

2. Dans les études d'encodage classiques, l'émotion exprimée est définie le plus souvent par l'intention d'un acteur et plus rarement par les caractéristiques d'une situation inductrice d'émotions. Etant donnée que tous les individus ne sont pas également expressifs et que la même situation n'induit pas toujours la même émotion chez différents individus, la reconnaissance, par un groupe d'auditeurs, des émotions simulées ou induites est parfois utilisée comme critère alternatif de l'émotion exprimée et comme garant de la validité écologique des expressions. Cette utilisation de la reconnaissance est toutefois problématique. La reconnaissance est – au même titre que l'intention expressive de l'acteur – soumise à la "compétence" des individus interrogés en matière de communication émotionnelle et l'hypothèse sous-jacente selon laquelle des auditeurs reconnaîtraient des émotions authentiques lorsqu'elles sont exprimées n'est pas vérifiée. L'exclusion des émotions "mal reconnues" nuit à l'exploration des différences éventuelles entre propriétés vocales des émotions exprimées et propriétés vocales perçues comme émotionnelles.

Ces deux premiers points mettent en évidence la nécessité d'**intégrer les études d'encodage et de décodage**. Considérer simultanément l'ensemble du processus de communication permet d'évaluer, d'une part, jusqu'à quel point les intentions (ou les sentiments) des locuteurs se traduisent en expressions vocales différenciées et, d'autre part, jusqu'à quel point les jugements des auditeurs peuvent capturer des expressions vocales différenciées. En outre, les caractéristiques vocales des émotions exprimées ainsi que des émotions perçues peuvent être identifiées et les

similarités/divergences entre les caractéristiques correspondant aux émotions exprimées et aux émotions perçues peuvent être examinées.

3. Dans le cadre de la recherche sur la communication vocale des émotions, l'intérêt de conceptualiser précisément les réactions émotionnelles est généralement sous-estimé. Les émotions et les expressions qui leurs correspondent sont rarement définies avec précision dans la littérature. La manière dont les réactions émotionnelles et leurs expressions peuvent être conceptualisées par des auditeurs/juges lors d'études de jugement n'est généralement pas d'avantage prise en considération et le choix des procédures de jugements utilisées est rarement justifié bien qu'il contraigne largement les jugements effectués. Une implication importante de cette conception imprécise des expressions émotionnelles étudiées se reflète dans le problème de l'influence du niveau d'*activation* émotionnelle sur les résultats acoustiques et sur les jugements. Dans la majorité des études, le niveau d'activation et le type d'émotion sont confondus, ce qui ne permet pas d'évaluer l'existence d'une différenciation entre types d'émotions différentes, indépendamment de leur niveau d'activation.

Pour le progrès des connaissances dans le domaine de la communication vocale des émotions, il est donc essentiel de définir précisément la manière dont les expressions émotionnelles étudiées sont conceptualisées. En particulier, il est important de **prendre en compte la dimension d'activation sous-jacente à la réaction émotionnelle**. Seule l'utilisation d'expressions correspondant à différents types d'émotions avec (pour chaque type d'émotion) différents niveaux d'activation, permet d'évaluer la possibilité de différencier – sur le plan acoustique et sur le plan des émotions perçues – différents types d'émotions ayant le même niveau d'activation.

4. La possibilité d'identifier des caractéristiques vocales correspondant à différentes émotions exprimées et/ou perçues est limitée par le choix a priori des caractéristiques vocales mesurées. Une critique souvent formulée relativement aux études de production (encodage) s'adresse à l'utilisation quasi-exclusive de mesures résumées sous forme de moyennes décrivant des segments de parole relativement longs (généralement une phrase grammaticale). La moyenne et/ou l'écart-type des contours de F0 et des contours d'intensité sont les mesures les plus fréquemment examinées. Une large part de l'information relative à l'*intonation* est éliminée par cette opération, bien que des arguments théoriques et quelques résultats empiriques indiquent que l'intonation (les contours de hauteur) pourrait jouer un rôle important dans la communication vocale des émotions.

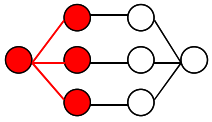
Pour parvenir à une différencier les émotions, il est nécessaire de mesurer un nombre suffisamment important de caractéristiques vocales différenciées. Dans ce domaine, il est important de développer de nouvelles mesures qui permettent de **représenter des caractéristiques vocales théoriquement**

importantes pour la communication émotionnelle telles que l'intonation. Il semble en particulier nécessaire d'analyser/décrire les contours de la F0 plutôt que de réduire ces contours à des indices de tendance centrale ou de variabilité globale.

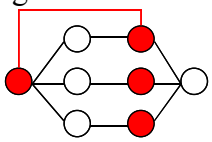
1.6 Objectifs et structure de la thèse

L'objectif principal des études décrites dans ce qui suit est d'examiner le *processus de communication* des émotions. En particulier, nous voulons évaluer dans quelle mesure un certain nombre de caractéristiques vocales sont capables de différencier différents types d'émotions exprimées et dans quelle mesure les mêmes caractéristiques vocales permettent de rendre compte de la reconnaissance des émotions exprimées. A cette fin un paradigme inspiré du *modèle en lentille* de Brunswik a été utilisé. Ce modèle (proposé par Scherer, 1978) distingue quatre étapes dans le processus de communication : premièrement l'état interne – dans le cas présent l'émotion – du locuteur ; deuxièmement l'extériorisation de cet état – c'est à dire ici les caractéristiques acoustiques de la voix qui encodent l'émotion du locuteur ; troisièmement la perception de ces caractéristiques vocales par un auditeur; quatrièmement l'inférence de l'état émotionnel du locuteur à partir des caractéristiques vocales perçues. Les sections 2 à 5 de cette thèse sont également définies dans le cadre des différentes étapes distinguées par ce paradigme. Les thèmes généraux abordés dans ces sections sont évoqués ci-dessous. Les thèmes et les problématiques spécifiques à chaque section seront présenté/es en détails au début de chaque section.

- Dans la **section 2**, les propriétés acoustiques d'un ensemble d'expressions émotionnelles sont examinées. Cette section permet d'identifier dans quelle mesure les caractéristiques acoustiques mesurées parviennent à différencier les émotions exprimées et de définir quelles caractéristiques vocales sont susceptibles de correspondre aux différentes émotions exprimées.

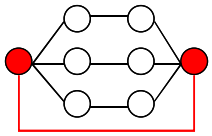


- Dans la **section 3**, un ensemble de caractéristiques vocales perçues (telles que la hauteur ou la rapidité perçues) sont évaluées dans une série d'études de jugements pour ces même expressions. Dans cette partie, les caractéristiques vocales perçues pour différentes émotions exprimées sont décrites. La possibilité de différencier les émotions exprimées par le biais des jugements relatifs aux caractéristiques vocales des expressions est évaluée.

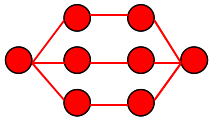


- Dans la **section 4**, une série d'études de jugements sont effectuées pour évaluer les émotions perçues dans les expressions étudiées. Cette partie permet d'établir la relation entre émotions

exprimées et émotions perçues. Les confusions et les différences sur le plan de la reconnaissance sont examinées pour les différentes émotions exprimées.



- Dans la **section 5**, les émotions exprimées, les caractéristiques acoustiques des expressions, les caractéristiques vocales perçues des expressions et les émotions attribuées aux expressions sont mises en relation. Le processus de communication est représenté dans son ensemble, ce qui permet d'évaluer le rôle et l'importance respective des différentes caractéristiques vocales – acoustiques et perçues – dans la communication vocale des émotions. Les différences éventuelles entre les caractéristiques vocales – acoustiques et perçues – correspondant aux émotions exprimées d'une part et aux émotions perçues d'autre part pourront être examinées dans cette partie.

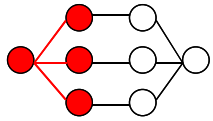


Une attention particulière a été accordée au choix des émotions étudiées, ainsi qu'à la procédure utilisée pour recueillir les attributions émotionnelles. Les expressions ont été sélectionnées dans une base de données de grande envergure incluant des expressions correspondant à 14 états émotionnels et produites par 12 acteurs. Dans cette base de données, les émotions exprimées sont clairement définies par des scénarios – décrivant des situations émotionnelles – fournis aux acteurs lors de l'enregistrement des expressions. Les expressions retenues ont été choisies de manière à inclure quatre *types* (ou *familles*) d'émotions et également deux niveaux d'*activation* par type d'émotion, soit au total huit états émotionnels. Cette approche a été définie afin d'évaluer si différentes caractéristiques vocales (acoustiques et perçues) parviennent à différencier non seulement *différents niveaux d'activation émotionnelle*, mais également *différents types d'émotions* exprimées. Les jugements émotionnels sont quant à eux réalisés en demandant aux auditeurs d'indiquer l'intensité de chaque type d'émotion contenue dans chaque expression. Cette approche a été choisie dans le but de contrôler un certain nombre de problèmes associés à la discrimination des émotions exprimées.

Finalement, la question de la contribution de l'*intonation* à la différenciation des émotions exprimées et à l'attribution d'émotions perçues a été examinée. De nombreuses propositions relativement au rôle de l'intonation dans la communication émotionnelle ont été formulées dans le passé. En particulier, il a été suggéré que la difficulté à distinguer différents types d'émotions exprimées sur le plan acoustique pourrait être expliquée par l'utilisation de mesures suprasegmentales moyennées à long terme (e.g. F0 moyenne, intensité moyenne) qui suppriment les informations contenues dans les *contours* de hauteur ou d'intensité. La dernière section de cette thèse à été consacrée à cette problématique:

- Dans la **section 6**, le codage d'un aspect important de l'intonation des expressions émotionnelles a été réalisé sous la forme d'une stylisation du contour de fréquence fondamentale. Ce codage de la courbe de F0 permet de tester la possibilité d'améliorer la différenciation entre les émotions exprimées en ajoutant une description du contour de hauteur aux mesures suprasegmentales plus traditionnelles (moyenne et écart-type/étendue de F0). D'autre part, différents contours de F0 ont été appliqués à des expressions de synthèse (text-to-speech). Des jugements relatifs à la qualité émotionnelle ont ensuite été obtenus pour ces expressions, afin d'évaluer l'influence de différentes caractéristiques des contours de F0 sur l'attribution émotionnelle.

2 Caractéristiques acoustiques des émotions exprimées



2.1 Introduction

L'objectif de cette première étape de l'analyse définie par le paradigme du modèle en lentille de Brunswik (Scherer, 1978) est de différencier sur le plan acoustique un ensemble d'expressions vocales correspondant à différentes émotions exprimées. De nombreuses études ayant été effectuées dans le passé afin d'identifier des "profils acoustiques" correspondant à différentes émotions, les principales difficultés liées à cette approche sont aujourd'hui bien identifiées. Dans une certaine mesure, une partie des problèmes rencontrés sont inhérents à la démarche même qui tente de faire correspondre un ensemble de caractéristiques acoustiques très spécifiques à des états émotionnels peu spécifiques. Toutefois, la reconnaissance explicite d'un certain nombre de difficultés rencontrées dans ce domaine de recherche permet de pallier à certaines d'entre elles et de définir plus précisément les objectifs ciblés.

Une limitation majeure des études visant à identifier des profils acoustiques correspondant à différentes émotions réside dans le caractère apparemment peu généralisable de leurs résultats. Les profils acoustiques identifiés dans le cadre d'études isolées ne sont souvent pas reproduits dans d'autres études. La faible réplication des résultats dans ce domaine a été attribuée à différents facteurs parmi lesquels le manque de précision dans la définition des états émotionnels étudiés ainsi que l'utilisation d'un nombre trop restreint de locuteurs et de contextes linguistiques semblent jouer un rôle prépondérant. S'agissant de définir très clairement les émotions exprimées, une stratégie qui s'est avérée efficace consiste à définir l'émotion par les caractéristiques d'une situation inductrice suffisamment prototypique pour ne pas laisser place à trop de variations interindividuelles sur le plan de la réaction et de l'expression induites par la situation. Les caractéristiques acoustiques des expressions ainsi définies ne devraient toutefois pas être considérées comme correspondant à un ensemble plus large de réactions pouvant être regroupées sous le même label (par exemple 'colère' ou 'joie') mais potentiellement différentes. De même les résultats obtenus pour un contexte linguistique donné ou pour un locuteur donné peuvent être, éventuellement, dus à une interaction entre ce contexte linguistique spécifique et l'émotion exprimée ou à une expression idiosyncrasique de ce locuteur particulier et ne devraient, en aucun cas, faire l'objet de généralisations à tous les autres contextes linguistiques et locuteurs possibles.

Un autre problème central concerne la part – apparemment assez faible – des résultats qui sont régulièrement répliqués d'une étude à l'autre. Depuis plusieurs décennies, les revues de la littérature effectuées dans ce domaine (par exemple: Frick, 1985; Johnstone & Scherer, 2000; Scherer, 1986)

constatent que les caractéristiques acoustiques décrites de manière répétée pour différentes émotions reflètent en fait essentiellement le degré d'activation émotionnelle. Les émotions faiblement activées (en particulier la "tristesse" qui est en général étudiée dans sa "version déprimée") se distinguent sur le plan acoustique des émotions fortement activées (en particulier la "colère" qui est en général étudiée dans sa "version chaude"), les profils établis pour les émotions avec une composante d'activation forte correspondant à des valeurs de fréquence fondamentale et d'intensité (moyennes, écart-types et/ou étendues) plus élevées et des valeurs de durées plus faibles (temps de pauses plus courts, vitesses d'élocution accrues) que pour les émotions avec une composante d'activation faible. Dans ce contexte, il apparaît que seul le croisement du niveau d'activation avec le type émotionnel (ou la famille émotionnelle) permettrait d'évaluer la possibilité d'établir des profils acoustiques qui reflèteraient non seulement l'activation émotionnelle mais également le type (ou la famille) émotionnel. A ce jour, deux tentatives ont été réalisées dans ce sens par Banse & Scherer (1996) et Juslin & Laukka (2001). Banse & Scherer parviennent à différencier les 14 types[3] d'émotions qu'ils étudient sur le plan acoustique, mais ils ne rapportent pas de résultats qui distinguent la contribution spécifique du niveau d'activation de celle du type d'émotion dans la discrimination réalisée. Ils soulignent cependant que, dans les classifications statistiques basées sur les mesures acoustiques, les émotions avec un niveau d'activation similaire, relativement aux émotions correspondant à différents niveaux d'activation, sont plus souvent "confondues" entre elles. Juslin & Laukka (2001) s'intéressent en fait à l'intensité émotionnelle qui, bien que liée à l'activation émotionnelle, s'en distingue partiellement (v. section 1.1.1.). Ils rapportent des effets importants de l'intensité émotionnelle et du type émotionnel sur la majorité des mesures acoustiques qu'ils ont effectuées. Ils notent d'une part que certains paramètres acoustiques sont plus affectés par l'intensité que par le type d'émotion et d'autre part que, pour d'autres paramètres, l'effet de l'intensité diffère en fonction du type d'émotion. Juslin & Laukka considèrent l'intensité comme une source de variance habituellement non-contrôlée dans les études qui visent à établir des profils acoustiques pour différentes émotions, ils attribuent à cette absence de contrôle du niveau d'intensité des émotions étudiées une partie de la difficulté, évoquée ci-dessus, à répliquer les profils acoustiques dans des études indépendantes. Ces auteurs insistent par ailleurs sur la nécessité de mesurer un nombre important de paramètres acoustiques dans le but d'identifier des profils acoustiques uniques pour différentes émotions.

[3] Parmi ces 14 types d'émotions, on trouve 4 paires d'émotions dont le premier membre correspond à une version faiblement active et le second membre à une version fortement active de la même émotion.

Face aux difficultés exposées ci-dessus, cette étude d'encodage se propose d'évaluer la possibilité de différencier sur le plan acoustique des expressions vocales correspondant à quatre types d'émotion (colère, joie, tristesse et peur) croisés avec deux niveaux d'activation (faible et forte). La possibilité de différencier les 8 catégories émotionnelles ainsi définies (colère froide, colère chaude, joie calme, joie exaltée, tristesse déprimée, tristesse désespérée, peur anxieuse, peur panique) et la contribution de l'activation à la différenciation effectuée seront évaluées. L'accent sera mis sur la différenciation des émotions exprimées plutôt que sur l'établissement de profils acoustiques qui tendent à être abusivement généralisés à des expressions qui correspondent à des états différents pouvant être regroupés sous le même label. Les expressions utilisées ont été extraites d'une base de donnée regroupant un ensemble plus large d'expressions produites par des acteurs professionnels. Elles sont définies clairement à l'aide de scénarios décrivant des situations inductrices des états émotionnels exprimés. Neuf acteurs et deux séquences de syllabes sans contenu linguistique (soit 18 expressions) ont été sélectionnés pour représenter chaque émotion afin de minimiser l'influence du contexte linguistique et des particularités expressives des locuteurs sur les résultats obtenus. Un grand nombre de paramètres acoustiques ont été mesurés afin d'accroître la possibilité de différencier un nombre relativement important d'expressions émotionnelles (8 catégories).

2.2 Méthode

2.2.1 Expressions émotionnelles extraites de la base de données de Munich

Les enregistrements utilisés dans cette étude sont extraits d'un ensemble d'expressions émotionnelles produites par des acteurs professionnels et enregistrées à Munich dans le cadre d'un projet de recherche dirigé par Klaus Scherer, Harald Wallbott, Rainer Banse et Heiner Ellgring. Douze acteurs allemands (six hommes et six femmes) ont été filmés et leurs expressions vocales ont été enregistrées alors qu'ils exprimaient quatorze types d'émotions définies par des scénarios. Chaque type d'émotion est représenté par deux scénarios distincts, les émotions simulées par les acteurs sont en conséquence définies par 28 situations émotionnelles. Les acteurs expriment les émotions correspondant à chaque scénario/situation en prononçant deux "phrases" constituées chacune de sept syllabes dépourvues de signification. Chaque "phrase" a, de plus, été répétée à deux reprises par chaque acteur et pour chaque scénario. Les scénarios fournis aux acteurs afin de définir les émotions qu'ils devaient exprimer, ainsi que les instructions qu'ils ont reçues, sont reproduits dans leur version originale allemande en annexe (annexes A1 et A2). Au total, 1344 expressions ont ainsi été enregistrées. En 1996, Banse & Scherer ont publié une étude incluant une sélection de 224 expressions vocales issues de cet ensemble d'enregistrements. Ces auteurs ont sélectionné pour chaque situation émotionnelle et chaque "phrase", les deux "meilleures" expressions produites par

les six acteurs et les deux "meilleures" expressions produites par les six actrices. Les "meilleurs" enregistrements pour les différentes émotions ont été définis relativement à l'authenticité des expressions émotionnelles et à la possibilité de reconnaître les émotions simulées par les acteurs. Cette qualité des expressions étant variable d'un acteur à l'autre, un nombre inégal d'expressions a été conservé pour chaque acteur. Des références plus détaillées concernant la constitution et la composition de cette base de donnée peuvent être consultées dans Banse & Scherer (1996).

Dans le cadre de l'étude présentée ci-dessous, 144 expressions ont été choisies parmi les 1344 expressions enregistrées. La sélection s'est basée sur les critères suivants:

- Les expressions sélectionnées ont été produites par neuf acteurs, 4 hommes et 5 femmes.
- Elles correspondent à huit types d'émotions exprimées.
- Pour chaque acteur et pour chaque type d'émotion une occurrence de chaque "phrase" a été sélectionnée aléatoirement.

Trois des douze acteurs ainsi que six des quatorze émotions de la base de données de Munich ont été éliminés de la sélection afin de réduire le nombre total des enregistrements utilisés. Les différents scénarios et les répétitions – n'ayant pas affecté les résultats publiés par Banse & Scherer (1996) – n'ont pas été pris en compte lors de la sélection.

2.2.1.1 Acteurs éliminés

Les acteurs masculins numéro 1 et numéros 6 ont été éliminés d'emblée sur la base des sélections effectuées préalablement par Banse & Scherer (1996). Bien que la locutrice numéro 10 soit identifiée par ces mêmes auteurs comme l'actrice la moins expressive, l'actrice numéro 5 (identifiée comme seconde actrice moins expressive par Banse & Scherer) a été éliminée en raison de la moindre qualité acoustique de ses enregistrements.

2.2.1.2 Emotions sélectionnées

Parmi les 14 émotions de la base de données de Munich, les 8 émotions suivantes ont été retenues[4]:

- la colère 'froide' (Kalter Ärger) *irritation* (*irrit.)*
- la colère 'chaude' (Heisser Ärger) *rage*
- la joie 'calme' (Stille Freude) *joie*

[4] Les labels retenus pour désigner chaque type d'émotion dans la suite du texte sont présentés en italiques ci-dessous. Les labels ont été choisis dans le but de fournir un descriptif court et qui permette de rendre compte à la fois du type d'émotion et du niveau d'activation. Ils sont définis dans le cadre de cette thèse comme correspondant aux labels allemands (entre parenthèse) soumis aux acteurs.

- la joie 'exaltée' (Überschäumende Freude) *exaltation (exalt.)*
- la tristesse 'déprimée' (Stille Trauer) *tristesse (trist.)*
- la tristesse 'désespérée' (Verzweiflung) *désespoir (désp.)*
- la peur 'anxieuse' (Angst) *anxiété (anx.)*
- la peur 'panique' (Panische Furcht) *panique (paniq*.)

Ce choix limite les émotions étudiées aux 4 types d'émotions les plus fréquemment examinées (colère, joie, tristesse, peur), pour lesquels il existe de nombreux résultats dans la littérature sur l'expression vocale des émotions. Les versions avec forte activation (colère 'chaude', joie 'exaltée', tristesse 'désespérée', peur 'panique') et avec faible activation (colère 'froide', joie 'calme', tristesse 'déprimée', peur 'anxieuse') de ces 4 types d'émotions ont été conservés pour les raisons suivantes:

- Afin d'évaluer, sur le plan acoustique, la possibilité de différencier non seulement le niveau d'activation, mais également les différents types d'émotions.
- Afin d'éviter, lors des études de jugement, une discrimination trop aisée des émotions encodées qui pourrait être basée essentiellement sur l'activation. L'identification de la tristesse serait en particulier facilitée dans la configuration la plus classique où les seuls enregistrements représentant une émotion peu active seraient les enregistrements de tristesse.

2.2.1.3 Description des expressions sélectionnées

La liste des 144 expressions sélectionnées figure dans le tableau 1. Les codes des expressions sont définis de la manière suivante:

- La première lettre correspond à l'émotion exprimée par l'acteur ; avec la valeur A pour 'peur anxieuse', F pour 'joie calme', H pour 'colère chaude', K pour 'colère froide', P pour 'peur panique', T pour 'tristesse déprimée', U pour 'joie exaltée', Z pour 'tristesse désespérée'.
- Le premier nombre – à la 2ème position dans le code de l'expression – correspond à l'un des deux scénarios fournis aux acteurs ; avec la valeur 1 pour le premier scénario et 2 pour le deuxième scénario (pour la description des scénarios, v. annexe A2).
- Le deuxième nombre correspond à la phrase prononcée par l'acteur. La valeur 1 correspond à la phrase "hätt sandik prong nju ventsie" et la valeur 2 à la phrase "fi gött leich jean kill gos terr".

- Le troisième nombre correspond à la répétition, chaque acteur ayant été enregistré à deux reprises avec le même scénario et la même phrase ; la valeur 1 correspond au premier enregistrement et la valeur 2 au deuxième enregistrement.
- Le quatrième nombre correspond au code d'identification de l'acteur ; les codes 02, 07, 09, 10 et 11 représentent les actrices sélectionnées, les codes 03, 04, 08 et 12 représentent les acteurs sélectionnés.

Les expressions vocales ont été enregistrées sur un support magnétique à l'aide d'un microphone de bonne qualité et d'un enregistreur professionnel "reel to reel" (REVOX). La distance entre les acteurs et le microphone a été maintenue constante. L'amplification du signal d'entrée a été ajustée pour chaque acteur, mais gardée constante pour toutes les expressions produites par un même acteur. Les enregistrements ont été digitalisés avec une fréquence d'échantillonnage de 44'100 Hz l'aide d'un convertisseur A/D (TANGO$_{24}$ de Frontier Design Group) et d'une carte son (DAKOTA de Frontier Design Group) de très bonne qualité.

Tableau 1: Liste des expressions vocales sélectionnées

émotion	phrase	actrice no2	acteur no3	acteur no4	actrice no7	acteur no8	actrice no9	actrice no10	actrice no11	acteur no12
Anx. (Angst)	1	A21202	A11103	A21204	A21207	A21108	A21209	A21210	A11111	A11112
	2	A22202	A22103	A22204	A22207	A22108	A22209	A22110	A12111	A12112
Joie (Stille Freude)	1	F21102	F21203	F11104	F11107	F11208	F21209	F21210	F11211	F21112
	2	F12202	F22103	F22104	F12107	F12208	F12209	F12210	F22211	F12112
Rage (Heisser Ärger)	1	H21202	H21203	H21204	H21207	H11108	H11209	H21210	H21111	H21212
	2	H22202	H22203	H22204	H22107	H12108	H22109	H22110	H12211	H12112
Irrit. (Kalter Ärger)	1	K21202	K11103	K21204	K11107	K21108	K11109	K21210	K11111	K11212
	2	K12202	K12203	K22204	K22107	K12108	K12209	K22210	K22111	K22212
Paniq. (Pani. Furcht)	1	P21202	P11103	P21204	P21107	P21208	P21209	P11210	P21111	P11212
	2	P12202	P12203	P22104	P12107	P22108	P22109	P22110	P12111	P12212
Trist (Stille Trauer)	1	T11102	T21203	T21204	T11207	T11108	T21209	T21210	T11111	T11212
	2	T22202	T22203	T12104	T12207	T12208	T22209	T22210	T22111	T22112
Exalt. (Über. Freude)	1	U21102	U21103	U11204	U11107	U11208	U11209	U11110	U11211	U21212
	2	U22102	U22103	U12204	U12207	U12208	U22209	U22210	U12111	U22212
Désp. (Verzweiflung)	1	Z11202	Z11203	Z21204	Z11207	Z21108	Z21109	Z21210	Z21111	Z11112
	2	Z22202	Z12203	Z12204	Z12107	Z22108	Z12209	Z22210	Z22211	Z12212

2.2.2 Analyses acoustiques

Des analyses acoustiques ont été effectuées en plusieurs étapes sur les 144 expressions émotionnelles sélectionnées.

2.2.2.1 Extraction de la fréquence fondamentale

Dans un premier temps, la fréquence fondamentale (F0) a été extraite par auto-corrélation à l'aide du logiciel d'analyse acoustique PRAAT développé par Paul Boersma (Boersma & Weenink, 1996). Une correction manuelle "conservatrice" du contour de F0 a ensuite été réalisée. Les erreurs de détection de F0 ont été corrigées uniquement lorsque l'algorithme a détecté un voisement (une valeur de F0) dans des parties clairement non-voisées du signal, c'est-à-dire sur des parties du signal correspondant à des consonnes non-voisées et sur des pauses. Dans de très rares cas, des sauts d'octave ont été corrigés en ajustant le niveau de la F0 détectée.

2.2.2.2 Segmentation

Les enregistrements ont ensuite été segmentés manuellement afin d'identifier le commencement et la fin exacts de chaque expression verbale, les parties correspondant à des segments de paroles et les "silences", les parties voisées et les parties non-voisées. Les segments identifiés comme "silencieux" contiennent en fait du son, mais jamais de la parole. Ils recouvrent les pauses véritables, les reprises de souffle et, dans quelques cas, des sons émis par les locuteurs, tels que des clics ou des bruits de respiration.

Cette segmentation a été réalisée par étapes, à l'aide de l'interface de segmentation disponible dans le logiciel PRAAT. Des segments phonétiques ont d'abord été identifiés en écoutant les enregistrements et en posant des frontières approximatives aux points de transitions entre les unités phonétiques. Ces frontières ont été ensuite ajustées en référence à la fois à la représentation graphique de l'onde sonore et aux phonèmes perceptibles dans chaque segment. Les frontières entre phonèmes étant difficiles à établir, des groupes de phonèmes ont été identifiés. Malgré l'utilisation de séquences de syllabes standards, il n'a pas été possible d'identifier les mêmes groupes de phonèmes dans toutes les expressions. Les réalisations varient en particulier pour des séquences telles que "prong niu" qui peut être réalisé avec ou sans la consonne finale de la première syllabe ; soit en séparant les deux syllabes ("prong … niu") ou en contractant les deux syllabes ("pro-niu"). La séquence "leich jean" est également réalisée de multiples façons, très peu d'occurrence de la syllabe "jean" sont réalisée avec la prononciation attendue (3ã, comme le prénom), les acteurs allemands remplacent cette syllabes par différentes réalisations telles que "schonn" ou "schann". Une contraction avec la syllabe précédente est dès lors fréquemment réalisée ("lei-schonn"). Les séquences réalisées n'étant pas parfaitement identiques sur le plan de leur contenu phonétique, la comparaison des résultats sur des segments courts (au niveau phonétique ou syllabique) pose problème ; des différences entre les expressions à ce niveau pourraient être dues à ces légères variations phonétiques qui peuvent être ou non la conséquence du type d'émotion exprimée. La

segmentation utilisée pour les analyses acoustiques et les comparaisons statistiques se limite donc à une segmentation regroupant les parties des expressions qui ne contiennent pas de parole ("silences"), les parties voisées et les parties non-voisées.

La distinction entre les parties voisées et non-voisées est basée sur la présence d'un voisement effectif, défini en écoutant les enregistrements et en regardant la représentation graphique de l'onde sonore. Dans cette perspective, certaines voyelles "soufflées" peuvent être considérées comme non-voisées. La segmentation des expressions en parties voisées et non-voisées a été ensuite mise en parallèle avec la courbe de F0. Lorsqu'un désaccord entre la segmentation manuelle et le tracé de F0 sont apparus, une correction a été effectuée. Pour les détections de F0 correspondant à des segments définis comme non-voisés, les points de F0 détectés ont été supprimés lorsque la source de l'erreur se situait clairement au niveau de la détection de F0 et les segments identifiés comme non-voisés ont été redéfinis comme voisés lorsque la source de l'erreur se situait clairement au niveau de la segmentation. En cas d'absence de détection de F0 dans des segments identifiés comme voisés, les segments ont été redéfinis comme non-voisés lorsque la source de l'erreur se situait manifestement au niveau de la segmentation; la détection de F0 n'a en revanche pas été modifiée lorsque la segmentation a été jugée correcte et l'absence de détection de la F0 erronée.

2.2.2.3 Valeurs extraites du contour de F0

Les valeurs suivantes ont été calculées sur la base du contour corrigé de F0 : la moyenne et l'écart-type ; les $5^{ème}$, $25^{ème}$, $50^{ème}$, $75^{ème}$ et $95^{ème}$ centiles ; le minimum, le maximum et l'étendue (différence entre le maximum et le minimum).

2.2.2.4 Valeurs extraites pour la durée

La durée totale des expressions a été mesurée en secondes. La durée relative des "silences", des segments voisés et non-voisés a été calculée en divisant la durée en secondes des différents segments par la durée totale des expressions. La durée des segments voisés relative uniquement aux segments articulés (voisé et non voisés, à l'exclusion des "silences") a été calculée en divisant la durée en secondes des segments voisés par la durée cumulée des segments voisés et non-voisés.

2.2.2.5 Valeurs extraites du contour d'intensité

Un contour d'intensité en dB a été calculé pour chaque expression à l'aide du logiciel PRAAT. La moyenne et l'écart-type ; le minimum, le maximum et l'étendue des valeurs d'intensité ont été extraits pour chaque expression. Lors des enregistrements une pré-amplification différente a été utilisée pour chaque locuteur, les valeurs d'intensité ne présentent donc pas une caractéristique de

mesure absolue et ne peuvent pas être comparées directement pour différents locuteurs. Tous les locuteurs ayant exprimé chaque émotion à deux reprises, les différences d'intensité entre les émotions exprimées par l'ensemble des locuteurs sont en revanche comparables.

2.2.2.6 Valeurs extraites pour la distribution de l'énergie dans le spectre

Différents indicateurs de la distribution de l'énergie dans le spectre ont été calculés séparément pour les parties voisées et les parties non-voisées. Les paramètres spectraux extraits par Banse & Scherer dans leur étude de 1996 ont été mesurés à l'aide du logiciel PRAAT.

Pour les parties voisées de chaque expression, les paramètres suivants ont été calculés : la proportion d'énergie dans les basses fréquences, l'indice de Hammarberg et la proportion d'énergie dans les bandes spectrales définies par Banse & Scherer (1996). La proportion d'énergie contenue dans les basses fréquences est indexée par deux mesures obtenues en divisant l'énergie contenue entre 0 et 500 Hz par l'énergie comprise entre 0 et 8000 Hz pour la première mesure et l'énergie contenue entre 0 et 1000 Hz par l'énergie contenue entre 0 et 8000 Hz pour la deuxième mesure. La proportion d'énergie comprise dans chacune des bandes définies par Banse & Scherer (125 à 200 Hz, 200 à 300 Hz, 300 à 500 Hz, 500 à 600 Hz, 600 à 800 Hz, 800 à 1000Hz, 1000 à 1600 Hz, 1600 à 5000 Hz, 5000 à 8000 Hz) a été calculée relativement à l'énergie comprise entre 0 et 8000 Hz. L'énergie comprise dans les bandes allant de 0 à 500 Hz, de 0 à 1000 Hz et de 0 à 8000 Hz, ainsi que l'énergie comprise dans les bandes définies par Banse & Scherer ont été extraites à l'aide du logiciel PRAAT. L'indice de Hammarberg a été calculé en soustrayant l'intensité maximale comprise entre 2000 et 5000 Hz de l'intensité maximale comprise entre 0 et 2000 Hz ; les maxima d'intensité ont été extraits des 2 bandes filtrées à l'aide du logiciel PRAAT.

Pour les parties non-voisées, la proportion d'énergie contenue dans les bandes spectrales définies par Banse & Scherer (1996) pour les segments non-voisés a été mesurée. La quantité d'énergie comprise dans chaque bande – soit entre 125 et 250 Hz, entre 250 et 400 Hz, 400 et 500 Hz, 500 et 1000 Hz, 1000 et 1600 Hz, 1600 et 2500 Hz, 2500 et 4000 Hz, 4000 et 5000 Hz et entre 5000 et 8000 Hz – a été extraite à l'aide du programme PRAAT, puis divisée par l'énergie comprise entre 0 et 8000 Hz. Deux mesures de la proportion d'énergie contenue dans les basses fréquences des parties non-voisées ont été obtenues en divisant l'énergie contenue entre 0 et 500 Hz et entre 0 et 1000 Hz par l'énergie compris entre 0 et 8000 Hz.

2.2.2.7 Codes utilisés pour désigner les paramètres acoustiques

Afin d'alléger la présentation des résultats, un code a été attribué à chacun des 44 paramètres acoustiques décrits ci-dessus. La liste des codes et leurs descriptifs sont présentés dans le tableau 2.

Tableau 2 : Codes et descriptifs des paramètres acoustiques

Code	Description	Code	Description
int.min	intensité minimale, mesurée en dB	**Pour les segments voisés**	
int.max	intensité maximale, mesurée en dB	v.125-200	% d'énergie entre 125 et 200 Hz
int.étdu	(Int_max-Int_min), étendue de l'intensité (dB)	v.200-300	% d'énergie entre 200 et 300 Hz
int.moy	intensité moyenne, mesurée en dB	v.300-500	% d'énergie entre 300 et 500 Hz
int.sd	écart-type de l'intensité, mesuré en dB	v.500-600	% d'énergie entre 500 et 600 Hz
		v.600-800	% d'énergie entre 600 et 800 Hz
F0.min	minimum de la fréquence fondamentale (Hz)	v.800-1k	% d'énergie entre 800 et 1000 Hz
F0.max	maximum de la fréquence fondamentale (Hz)	v.1k-1.6k	% d'énergie entre 1000 et 1600 Hz
F0.étdu	(F0_max-F0_min), étendue de la F0 (Hz)	v.1.6k-5k	% d'énergie entre 1600 et 5000 Hz
F0.moy	moyenne de la fréquence fondamentale (Hz)	v.5k-8k	% d'énergie entre 5000 et 8000 Hz
F0.sd	écart-type de la fréquence fondamentale (Hz)	v.0-500	% d'énergie en dessous de 500 Hz
F0.05c	5ème centile de la fréquence fondamentale (Hz)	v.0-1k	% d'énergie en dessous de 1000 Hz
F0.25c	25ème centile de la fréquence fondamentale (Hz)		
F0.50c	médiane de la fréquence fondamentale (Hz)	Hamm	indice de Hammarberg, intensité max. entre 0 et 2 kHz moins intensité max. entre 2 et 5 kHz
F0.75c	75ème centile de la fréquence fondamentale (Hz)		
F0.95c	95ème centile de la fréquence fondamentale (Hz)	**Pour les segments non-voisés**	
F0.05-95c	(F0_95c-F0_05c), étendue de la F0 (Hz)	n.125-250	% d'énergie entre 125 et 250 Hz
		n.250-400	% d'énergie entre 250 et 400 Hz
dur.s/tot	durée des segm. silencieux sur la durée totale	n.400-500	% d'énergie entre 400 et 500 Hz
dur.n/tot	durée des segm. non-voisés sur la durée totale	n.500-1k	% d'énergie entre 500 et 1000 Hz
dur.v/tot	durée des segm. voisés sur la durée totale	n.1k-1.6k	% d'énergie entre 1000 et 1600 Hz
dur.v/art	durée des segm. voisés sur la durée articulée	n.1.6k-2.5k	% d'énergie entre 1600 et 2500 Hz
dur.tot	durée totale des expressions (en secondes)	n.2.5k-4k	% d'énergie entre 2500 et 4000 Hz
		n.4k-5k	% d'énergie entre 4000 et 5000 Hz
		n.5k-8k	% d'énergie entre 5000 et 8000 Hz
		n.0-500	% d'énergie en dessous de 500 Hz
		n.0-1k	% d'énergie en dessous de 1000 Hz

2.3 Résultats

2.3.1 Fiabilité des mesures acoustiques

Les analyses décrites ci-dessus ont été partiellement effectuées manuellement ; la segmentation – en silences et parties voisées/non-voisée – a été réalisée à la main et le contour de F0 a également été corrigé manuellement. Relativement à des analyses entièrement automatisées, la segmentation manuelle et le contrôle de la détection de F0 permettent en principe de réduire un certain nombre d'erreurs qui peuvent survenir lors de l'analyse automatique. Dans le contexte de l'étude de la parole émotionnelle simulée par des acteurs des erreurs telles que la détection erronée de F0 (en l'absence de segment de parole voisé) ou de parole (en présence de bruit) sont particulièrement fréquentes en raison de la très grande variabilité des productions vocales. En revanche, la segmentation et la correction manuelle sont susceptibles d'introduire d'autres erreurs liées aux jugements effectués et aux décisions prises par le correcteur humain. Elles présentent donc le désavantage de n'être pas parfaitement reproductibles. Afin de contrôler au moins partiellement la validité des mesures

acoustiques effectuées, une partie des analyses ont été reproduites à l'aide d'un programme d'analyse entièrement automatique développé à la faculté de psychologie de l'Université de Genève dans le cadre d'un projet de recherche financé par le FNRS. L'extraction entièrement automatique d'une partie des paramètres liés à la F0 et d'une partie des paramètres liés à la distribution de l'énergie dans le spectre a été comparée aux calculs effectués et utilisés dans cette étude. Le tableau 3 présente les corrélations entre les mesures effectuées à l'aide du programme d'analyses acoustiques automatiques et les mesures corrigées manuellement. Ces mesures qui intègrent chacune une part d'erreur – liée pour les unes aux procédures automatiques et pour les autres à l'attention et au jugement humain – restent cependant très fortement corrélées et révèlent une fiabilité satisfaisante des mesures comparées.

Tableau 3 : Fiabilité des mesures acoustiques. Corrélations entre une partie des mesures acoustiques effectuées avec segmentation/correction manuelle et des mesures entièrement automatiques.

Mesure	Corrélation		Mesure	Corrélation	
	r	N*		r	N*
dur.tot	0.90	141	v.125-200	0.98	141
			v.200-300	0.98	141
F0.05c	0.86	141	v.300-500	0.99	141
F0.25c	0.99	141	v.500-600	0.98	141
F0.75c	0.97	141	v.600-800	0.99	141
F0.moy	0.97	141	v.800-1k	0.99	141
F0.sd	0.80	141	v.1k-1.6k	1.00	141
F0.max	0.89	94**	v.1.6k-5k	0.94	141
			v.5k-8k	0.97	141
			v.0-500	0.99	141
Hamm	0.84	141	v.0-1k	0.97	141

*l'analyse automatique a échoué pour 3 enregistrements
**les maxima de F0 dépassant un certain seuil ont été supprimés lors de l'analyse automatique

2.3.2 Relations entre les paramètres acoustiques et les émotions exprimées.

2.3.2.1 Standardisation des paramètres acoustiques pour chaque locuteur

Dans le domaine des caractéristiques vocales, une large partie de la variabilité des données est relative aux caractéristiques personnelles des locuteurs. Les valeurs ayant trait à la fréquence fondamentale sont, par exemple, sensibles non seulement au sexe du locuteur mais se situent également dans une étendue donnée pour chaque locuteur. Cet argument s'étend à l'ensemble des paramètres acoustiques mesurés et tout particulièrement aux valeurs relatives à l'intensité des expressions (la moyenne, l'écart type, le minimum et le maximum des valeurs d'intensité) qui dépendent également du niveau d'amplification du signal ajusté séparément pour chaque locuteur. Afin de tester les différences entre émotions exprimées, tous les paramètres acoustiques mesurés ont

donc été standardisés séparément pour chaque locuteur. Cette standardisation des résultats par locuteur permet de contrôler l'influence des différences systématiques entre les locuteurs qui, sans quoi, affecteraient une partie des résultats présentés ci-dessous. A titre indicatif, les moyennes et les écarts-types par émotion exprimée sont présentés en annexe (A3) pour tous les paramètres avant la standardisation.

2.3.2.2 Tests de l'effet des "phrases" et des émotions exprimées sur l'ensemble des paramètres acoustiques

Des ANOVAs à mesures répétées ont été effectuées afin de tester l'influence des "phrases" (2 séquences de syllabes dépourvues de contenu sémantique) et des émotions exprimées (8 catégories) sur chacun des paramètres acoustiques standardisés par locuteur. En raison de contraintes d'espace, les résultats de ces 44 ANOVAs sont présentés de manière fractionnée dans les tableaux 4 et 5. Le tableau 4 inclut les effets principaux de la phrase, de l'émotion et leurs interactions sur les 21 paramètres acoustiques dérivés du contour de fréquence fondamental, du contour d'intensité et de la durée, ainsi que sur l'indice de Hammarberg. Le tableau 5 représente les effets principaux de la phrase, de l'émotion et leurs interactions sur la proportion d'énergie contenue dans les 22 bandes spectrales extraites des segments voisés et non-voisés.

On observe un effet principal significatif de la phrase pour 12 paramètres acoustiques et un effet principal significatif de l'émotion exprimée pour 35 paramètres acoustiques. Les effets significatifs de la phrase semblent être légèrement plus forts que les effets significatifs de l'émotion (pour les effets significatifs de la phrase, η^2 moyen = .60 ; pour les effets significatifs de l'émotion, η^2 moyen = .55). Mais l'effet de l'émotion est en général plus important que l'effet de la phrase (pour la phrase, η^2 moyen = .26 ; pour l'émotion, η^2 moyen = .47). De manière prévisible, les effets principaux de la phrase apparaissent surtout pour la durée relative des segments voisés et non-voisés (dur.n/tot et dur.v/art) et pour la proportion d'énergie contenue dans certaines bandes spectrales (en particulier v.0-500, v.0-1k, n.2.5k-4k et n.4k-5k). Les émotions exprimées ont un effet significatif sur la plupart des paramètres acoustiques. Les résultats des ANOVAs indiquent que seules la durée relative des segments non-voisée (dur.n/tot) et quelques bandes spectrales, en particulier dans le domaine du spectre non-voisé, ne sont pas significativement affectées par l'émotion exprimée.

D'autre part, 6 effets d'interaction de l'émotion et de la phrase sont significatifs. En comparaison avec les effets principaux de l'émotion et de la phrase, ces effets sont relativement faibles : $\eta^2 = .36$ pour dur.v/art (durée relative des segments voisé par rapport aux segments articulés), les autres effets significatifs de l'interaction phrase-émotion s'observent pour une partie des bandes spectrales non-voisées, la taille de ces effets varie de $\eta^2 = .24$ (pour n.250-400) à $\eta^2 = .34$ (n.5k-8k).

Tableau 4: ANOVAs à mesures répétées - Effets "within" de la phrase et de l'émotion sur 22 paramètres acoustiques standardisés par locuteur (valeurs d'intensité, de F0, de durée et indice de Hammarberg), les résultats significatifs (p<.05) sont indiqués en gras.

Code	source	df	F	sig.	eta^2	Code	source	df	F	sig.	eta^2
int.min	phrase	(1,8)	0.64	.446	0.07	dur.s/tot	phrase	(1,8)	0.24	.635	0.03
	emo	**(7,56)**	**9.56**	**.000**	**0.54**		**emo**	**(7,56)**	**9.12**	**.000**	**0.53**
	phrase*emo	(7,56)	1.78	.110	0.18		phrase*emo	(7,56)	1.40	.224	0.15
int.max	phrase	(1,8)	0.00	.947	0.00	dur.n/tot	**phrase**	**(1,8)**	**20.35**	**.002**	**0.72**
	emo	**(7,56)**	**35.23**	**.000**	**0.81**		emo	(7,56)	1.18	.327	0.13
	phrase*emo	(7,56)	1.16	.338	0.13		phrase*emo[5]	(7,56)	2.99	.010	0.27
int.étdu	phrase	(1,8)	0.02	.890	0.00	dur.v/tot	phrase	(1,8)	3.54	.097	0.31
	emo	**(7,56)**	**7.73**	**.000**	**0.49**		**emo**	**(7,56)**	**12.37**	**.000**	**0.61**
	phrase*emo	(7,56)	0.56	.781	0.07		phrase*emo	(7,56)	1.99	.073	0.20
int.moy	phrase	(1,8)	0.06	.817	0.01	dur.v/art	**phrase**	**(1,8)**	**20.84**	**.002**	**0.72**
	emo	**(7,56)**	**35.82**	**.000**	**0.82**		**emo**	**(7,56)**	**5.27**	**.000**	**0.40**
	phrase*emo	(7,56)	1.36	.242	0.14		**phrase*emo**	**(7,56)**	**4.58**	**.000**	**0.36**
int.sd	phrase	(1,8)	0.46	.517	0.05	dur.tot	phrase	(1,8)	1.08	.329	0.12
	emo	**(7,56)**	**7.55**	**.000**	**0.49**		**emo**	**(7,56)**	**6.66**	**.000**	**0.45**
	phrase*emo	(7,56)	0.78	.603	0.09		phrase*emo	(7,56)	1.08	.389	0.12
F0.min	phrase	(1,8)	0.06	.820	0.01	F0.05c	phrase	(1,8)	0.00	.948	0.00
	emo	**(7,56)**	**14.10**	**.000**	**0.64**		**emo**	**(7,56)**	**18.18**	**.000**	**0.69**
	phrase*emo	(7,56)	1.37	.237	0.15		phrase*emo	(7,56)	0.84	.558	0.10
F0.max	phrase	(1,8)	0.85	.383	0.10	F0.25c	phrase	(1,8)	0.26	.623	0.03
	emo	**(7,56)**	**15.89**	**.000**	**0.67**		**emo**	**(7,56)**	**23.13**	**.000**	**0.74**
	phrase*emo	(7,56)	0.74	.641	0.08		phrase*emo	(7,56)	0.71	.667	0.08
F0.étdu	phrase	(1,8)	1.59	.243	0.17	F0.50c	phrase	(1,8)	0.67	.438	0.08
	emo	**(7,56)**	**8.98**	**.000**	**0.53**		**emo**	**(7,56)**	**19.45**	**.000**	**0.71**
	phrase*emo	(7,56)	1.11	.367	0.12		phrase*emo	(7,56)	0.92	.496	0.10
F0.moy	phrase	(1,8)	0.83	.390	0.09	F0.75c	phrase	(1,8)	0.93	.363	0.10
	emo	**(7,56)**	**23.10**	**.000**	**0.74**		**emo**	**(7,56)**	**21.00**	**.000**	**0.72**
	phrase*emo	(7,56)	0.95	.473	0.11		phrase*emo	(7,56)	1.11	.372	0.12
F0.sd	phrase	(1,8)	1.72	.226	0.18	F0.95c	phrase	(1,8)	1.65	.235	0.17
	emo	**(7,56)**	**7.08**	**.000**	**0.47**		**emo**	**(7,56)**	**16.44**	**.000**	**0.67**
	phrase*emo	(7,56)	1.38	.233	0.15		phrase*emo	(7,56)	1.07	.396	0.12
Hamm	phrase	(1,8)	4.51	.067	0.36	F0.05-95	phrase	(1,8)	2.06	.189	0.20
	emo	**(7,56)**	**14.70**	**.000**	**0.65**		**emo**	**(7,56)**	**8.27**	**.000**	**0.51**
	phrase*emo	(7,56)	1.00	.438	0.11		phrase*emo	(7,56)	1.27	.280	0.14

Dans le but d'examiner les résultats par émotion indépendamment des syllabes prononcées, les résultats obtenus pour les 2 "phrases" ont été regroupés pour la suite des analyses. Les effets significatifs des différentes émotions exprimées sur les paramètres acoustiques mesurés sont représentés avec plus de précisions, en partie ci-dessous dans la section *discrimination des émotions exprimées* (graphiques 1 à 8) et en totalité dans les annexes (annexe A4 et annexe A5).

[5] Cet effet devient non significatif avec la correction de Greenhouse-Geisser: F(2.45, 19.62) = 2.99, p = .065

Tableau 5: ANOVAs à mesures répétées - Effets "within" de la phrase et de l'émotion sur 22 paramètres acoustiques standardisés par locuteur (paramètres spectraux), les résultats significatifs (p<.05) sont indiqués en gras.

Code	source	df	F	sig.	eta^2	Code	source	df	F	sig.	eta^2
v.125-200	phrase	(1,8)	1.99	.196	0.20	n.125-250	phrase	(1,8)	1.63	.237	0.17
	emo	**(7,56)**	**11.52**	**.000**	**0.59**		emo	(7,56)	1.38	.232	0.15
	phrase*emo	(7,56)	1.30	.267	0.14		**phrase * emo**	**(7,56)**	**2.94**	**.011**	**0.27**
v.200-300	**phrase**	**(1,8)**	**5.44**	**.048**	**0.40**	n.250-400	phrase	(1,8)	0.76	.409	0.09
	emo	**(7,56)**	**4.04**	**.001**	**0.34**		**emo**	**(7,56)**	**3.83**	**.002**	**0.32**
	phrase*emo	(7,56)	1.38	.230	0.15		**phrase * emo**	**(7,56)**	**2.51**	**.026**	**0.24**
v.300-500	**phrase**	**(1,8)**	**6.25**	**.037**	**0.44**	n.400-500	phrase	(1,8)	0.93	.364	0.10
	emo	(7,56)	1.09	.383	0.12		emo	(7,56)	0.61	.747	0.07
	phrase*emo	(7,56)	1.89	.089	0.19		phrase * emo	(7,56)	1.21	.311	0.13
v.500-600	phrase	(1,8)	1.66	.234	0.17	n.500-1k	phrase	(1,8)	3.99	.081	0.33
	emo	(7,56)	1.81	.103	0.18		**emo**	**(7,56)**	**5.28**	**.000**	**0.40**
	phrase*emo	(7,56)	1.80	.106	0.18		phrase * emo	(7,56)	1.10	.375	0.12
v.600-800	**phrase**	**(1,8)**	**5.48**	**.047**	**0.41**	n.1k-1.6k	phrase	(1,8)	1.76	.221	0.18
	emo	**(7,56)**	**2.36**	**.034**	**0.23**		**emo**	**(7,56)**	**5.62**	**.000**	**0.41**
	phrase*emo	(7,56)	0.68	.691	0.08		phrase * emo	(7,56)	0.88	.527	0.10
v.800-1k	phrase	(1,8)	3.22	.111	0.29	n.1.6k-2.5k	phrase	(1,8)	0.82	.392	0.09
	emo	**(7,56)**	**5.48**	**.000**	**0.41**		**emo**	**(7,56)**	**5.92**	**.000**	**0.43**
	phrase*emo	(7,56)	1.32	.258	0.14		**phrase * emo**	**(7,56)**	**2.46**	**.028**	**0.24**
v.1k-1.6k	phrase	(1,8)	4.26	.073	0.35	n.2.5-4k	**phrase**	**(1,8)**	**57.47**	**.000**	**0.88**
	emo	**(7,56)**	**16.00**	**.000**	**0.67**		**emo**	**(7,56)**	**3.04**	**.009**	**0.28**
	phrase*emo	(7,56)	0.70	.676	0.08		**phrase * emo**	**(7,56)**	**3.25**	**.006**	**0.29**
v.1.6k-5k	**phrase**	**(1,8)**	**12.76**	**.007**	**0.61**	n.4k-5k	**phrase**	**(1,8)**	**42.16**	**.000**	**0.84**
	emo	**(7,56)**	**14.68**	**.000**	**0.65**		emo	(7,56)	1.09	.385	0.12
	phrase*emo	(7,56)	1.04	.414	0.12		phrase * emo	(7,56)	1.12	.366	0.12
v.5k-8k	phrase	(1,8)	1.84	.212	0.19	n.5k-8k	phrase	(1,8)	0.26	.621	0.03
	emo	**(7,56)**	**6.49**	**.000**	**0.45**		emo	(7,56)	0.60	.756	0.07
	phrase*emo	(7,56)	1.04	.412	0.12		**phrase * emo**	**(7,56)**	**4.07**	**.001**	**0.34**
v.0-500	**phrase**	**(1,8)**	**17.21**	**.003**	**0.68**	n.0-500	**phrase**	**(1,8)**	**5.67**	**.044**	**0.41**
	emo	**(7,56)**	**9.59**	**.000**	**0.55**		emo[6]	(7,56)	3.35	.005	0.30
	phrase*emo	(7,56)	0.47	.855	0.06		phrase * emo	(7,56)	1.80	.106	0.18
v.0-1k	**phrase**	**(1,8)**	**16.93**	**.003**	**0.68**	n0-1k	**phrase**	**(1,8)**	**6.39**	**.035**	**0.44**
	emo	**(7,56)**	**18.46**	**.000**	**0.70**		emo	(7,56)	1.36	.241	0.15
	phrase*emo	(7,56)	1.07	.393	0.12		phrase * emo	(7,56)	1.86	.095	0.19

2.3.2.3 Sélection des paramètres acoustiques utilisés pour la discrimination

Les paramètres acoustiques mesurés reflètent des caractéristiques vocales qui sont liées sur le plan de la production vocale, ils ne sont donc pas indépendants. Une augmentation de l'effort vocal, par exemple, produira une élévation de l'intensité et de la F0 ainsi que des modifications dans la distribution de l'énergie spectrale. De plus, certains paramètres correspondent à des indices légèrement différents du même aspect vocal : Le rang et l'écart-type de la F0 et de l'intensité sont 2 mesures de la variabilité de la F0 et de l'intensité, le maximum et le 95ème centile de F0 sont 2

[6] Cet effet devient non significatif avec la correction de Greenhouse-Geisser: F(2.17, 17.35) = 3.35, p = .056

mesures du "plafond" de F0, le minimum et le 5ème centile de F0 sont 2 mesures du "plancher" de F0. La proportion d'énergie en dessous de 500 et de 1000 Hz ainsi que l'indice de Hammarberg sont des indicateurs de la distribution de l'énergie entre les hautes et les basses fréquences du spectre. Les corrélations, en partie très élevées, entre les paramètres acoustiques sont rapportées dans trois tableaux en annexe (annexes A6.1 – A6.3). Afin de réduire cette forte colinéarité, il s'est avéré nécessaire d'éliminer une partie des variables acoustiques avant de procéder à d'autres analyses statistiques. A cette fin, une analyse en composantes principale a été effectuée sur les 44 paramètres acoustiques mesurés. Les résultats de cette analyse sont présentés dans le tableau 6 (valeurs propres et pourcentage de variance expliquée par les composantes de l'ACP avant et après rotation) et le tableau 7 (saturations des paramètres acoustiques avec les composantes dégagées par l'ACP).

Tableau 6 : Valeurs propres des composantes de l'ACP et pourcentage de variance expliquée par les composantes avant et après rotation (Varimax).

Composante	Avant rotation			Rotation Varimax		
	valeur propre	% de variance	% cumulé de variance	valeur propre	% de variance	% cumulé de variance
1	15.73	35.75	35.75	7.78	17.68	17.68
2	4.64	10.54	46.29	4.62	10.51	28.19
3	3.04	6.90	53.19	4.54	10.32	38.51
4	2.83	6.44	59.62	4.10	9.31	47.82
5	2.53	5.76	65.38	3.94	8.96	56.78
6	2.06	4.67	70.05	3.59	8.16	64.94
7	1.78	4.05	74.10	2.49	5.66	70.60
8	1.55	3.52	77.62	2.22	5.05	75.66
9	1.20	2.73	80.35	2.07	4.69	80.35

Après rotation, les 6 premières composantes de l'ACP sont facilement interprétables sur le plan acoustique/vocal (v. tableau 7). Dans le but de conserver pour la suite de la présentation des résultats des variables acoustiques plus directement interprétables et dont la mesure peut être répliquée, la sélection d'un paramètre pour représenter chaque composante a été préférée à l'utilisation des scores factoriels. Pour chacune des 9 composantes, un paramètre acoustique représentant bien la composante (i.e. avec une saturation forte sur la composante considérée et avec des saturations faibles sur les autres composantes) et la dimension théoriquement sous-jacente à la composante (i.e. le paramètre correspondant le mieux à l'interprétation de la composante) a été sélectionné. Les paramètres sélectionnés et les critères spécifiques utilisés sont décrits ci-dessous.

Tableau 7 : Saturations des paramètres acoustiques avec les composantes de l'ACP (rotation Varimax), les saturations inférieures à 0.3 ne sont pas représentées.

Paramètres	Composantes de l'ACP								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
F0.sd	0.970								
F0.05-95c	0.966								
F0.étdu	**0.944**								
F0.95c	0.856	0.400							
F0.max	0.849	0.397							
F0.75c	0.753	0.501							
F0.moy	0.674	0.616							
F0.50c	0.653	0.592	0.313						
F0.min		**0.911**							
F0.05c		0.876							
F0.25c	0.460	0.750							
v.1.6k-5k			0.860						
v.0-1k			**-0.839**						
Hamm	-0.316		-0.748						
v.1k-1.6k			0.621					0.335	
v.5k-8k		0.316	0.523					-0.469	
v.800-1k			0.423						
dur.tot				**-0.844**					
dur.s/tot				-0.823					
dur.v/tot				0.791			-0.324		
int.min	0.319			0.666					
int.moy	**0.479**	**0.335**	**0.397**	**0.548**					
n.0-1k					**0.908**				
n.0-500					0.764	-0.352			
n.5k-8k					-0.688		-0.473		
n.400-500					0.642				
n.4k-5k					-0.592				
n.500-1k	0.341			0.339	0.554				
n.125-250					0.551				
n.1k-1.6k	0.421				0.486				
n.250-400		0.364			0.462	0.333			
int.sd						0.852			
int.étdu						**0.847**			
v.125-200		-0.318				-0.531			
int.max	0.455	0.334	0.303	0.363		0.520			
dur.n/tot						-0.330	0.763		
n.2.5-4k					-0.307		0.699		
dur.v/art				0.532		0.345	**-0.635**		
n.1.6k-2.5k						0.347	0.611		
v.600-800								**0.785**	
v.200-300								-0.662	
v.300-500									**-0.867**
v.500-600									0.752
v.0-500			-0.378					-0.426	-0.672

La **première composante** rassemble principalement les paramètres qui décrivent la **variabilité de la fréquence fondamentale**. L'**étendue de F0** (F0.étdu) qui sature clairement sur la première composante représente le paramètre qui, en théorie, décrit le mieux la variabilité de la F0. L'étendue du 5ème au 95ème centile correspond à une estimation "atténuée" de l'étendue de F0 destinée en principe à corriger les éventuelles erreurs de détection de la F0 qui ont été, dans le cas présent, corrigées manuellement. L'écart-type de F0 inclut les variations à court et à long terme de la F0, il représente de ce fait un paramètre plus "hétérogène"; un écart-type important pourrait correspondre à des fluctuations de grande étendue du contour de F0 à long terme ou à un phénomène de fluctuations de faible étendue et à plus court terme.

La **deuxième composante** semble correspondre au **niveau le plus faible** atteint par la F0 dans les expressions analysées (la valeur plancher/floor de F0). Le **minimum de F0** (F0.min) qui présente la saturation la plus forte sur ce facteur est également la variable qui décrit le plus directement cette dimension vocale.

La **troisième composante** rassemble des paramètres qui correspondent à la **répartition de l'énergie entre les hautes et les basses fréquences du spectre voisé**. La mesure la plus courante de cet aspect est la **proportion d'énergie en dessous de 1000 Hz** (v.0-1k), c'est également l'une des deux variables qui présentent les plus fortes saturations sur cette composante.

La **quatrième composante** rassemble des mesures relatives à la **durée des expressions vocales**. La **durée totale** (dur.tot) qui présente la plus forte saturation sur cette composante a été choisie pour représenter cette composante.

La **cinquième composante** rassemble les paramètres acoustiques qui décrivent la **distribution de l'énergie dans le spectre non-voisé**. La **proportion d'énergie contenue dans les basses fréquences du spectre non-voisé** (n.0-1k) est la variable qui sature le plus fortement sur cette composante. C'est également la mesure la plus générale de la répartition de l'énergie non-voisé dans le spectre.

La **sixième composante** semble correspondre surtout à la **variabilité de l'intensité** des expressions. L'**étendue de l'intensité** (int.étdu) a été sélectionnée pour représenter cette composante. Cette variable a été préférée à l'écart-type des valeurs d'intensité qui est affecté à la fois par de grandes excursions (à long terme) du contour d'intensité et par de plus nombreuses petites excursions à court terme.

La **septième composante** est plus difficile à interpréter. Elle rassemble 2 paramètres relatifs à la distribution de l'énergie dans le spectre non-voisé (n.2.5k-4k, n.1.6k-2.5k) et 2 paramètres relatifs à la **durée relative des segments voisés/non voisés** (dur.v/art, dur.n/tot) ; ces 4 paramètres saturent également, dans une moindre mesure, sur une ou plusieurs autres composantes. Le **rapport entre la**

durée des segments voisés et la durée de segments non-voisés (dur.v/art) présente une saturation relativement élevée sur la composante 4 (représentée par la durée totale des expressions), mais il constitue une mesure théoriquement plus intéressante que les trois autres mesures, il a été sélectionné pour cette raison.

Les **huitième** et la **neuvième composantes** rassemblent **différentes bandes du spectre voisés**. Les 2 paramètres qui présentent les plus fortes saturations sur la composante 8 (**v.600-800**) d'une part et sur la composante 9 (**v.300-500**) d'autre part ont été sélectionnés pour représenter ces composantes.

En théorie, **l'intensité des expressions vocales** devrait être fortement affectée par l'émotion exprimée. Parmi les mesures effectuées sur les expressions utilisées dans cette étude, l'intensité moyenne (int.moy) et l'intensité maximale (int.max) sont les paramètres les plus affectés par le type d'émotion exprimée ; 82% de la variance résiduelle[7] de int.moy et 81% de la variance résiduelle de int.max dépendent du type d'émotion exprimée. De ce fait, et bien que ces paramètres saturent relativement faiblement et de manière comparable sur les 4 premières composantes de l'ACP (ainsi que sur la 6ème pour l'intensité maximale), l'**intensité moyenne** sera comprise dans la suite des analyses présentées ci-dessous. Les paramètres acoustiques retenus pour la suite de la description des résultats sont représentés en gras dans le tableau 7.

Les corrélations entre les paramètres acoustiques sélectionnés sont présentées dans le tableau 8. La moitié, environ, des corrélations représentées dans ce tableau sont significatives. Toutefois, les corrélations les plus importantes représentées dans les tableaux en annexe (v. annexes A6.1 – A6.3) ont été éliminées par la sélection des paramètres. De manière prévisible, les corrélations les plus importantes sont observées pour l'intensité moyenne qui saturait sur plusieurs composante de l'ACP et qui corrèle de manière significative avec tous les autres paramètres sélectionnés.

Tableau 8 : Corrélations entre les paramètres acoustiques sélectionnés, les corrélations significatives ($p<.05$) sont indiquées en gras.

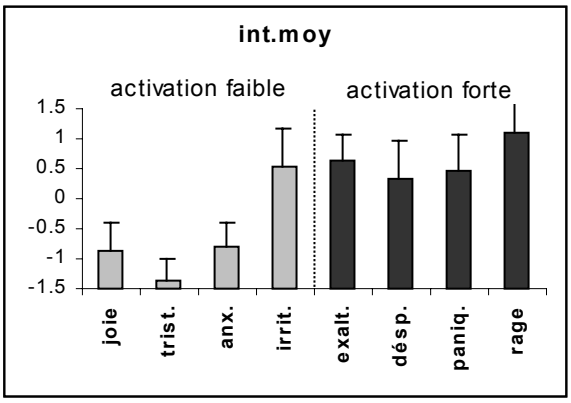
	int.étdu	int.moy	F0.min	F0.étdu	dur.tot	dur.v/art	v.0-1k	v.300-500	v.600-800
int.moy	**0.44**								
F0.min	**0.34**	**0.45**							
F0.étdu	**0.35**	**0.60**	0.13						
dur.tot	0.04	**-0.39**	-0.05	-0.03					
dur.v/art	**0.29**	**0.59**	0.10	**0.41**	**-0.33**				
v.0-1k	**-0.30**	**-0.73**	**-0.39**	**-0.42**	**0.27**	**-0.34**			
v.300-500	-0.08	**-0.17**	-0.11	0.03	0.03	0.02	**0.30**		
v.600-800	**0.23**	**0.29**	0.12	0.05	-0.02	0.14	-0.13	-0.15	
n.0-1k	**-0.26**	**-0.19**	-0.10	-0.02	0.11	-0.07	**0.20**	0.10	-0.11

[7] La part de variance expliquée par le 'locuteur', la 'phrase' et les interactions entre 'locuteur' 'phrase' et 'émotion' sont retirées préalablement.

2.3.2.4 Discrimination des émotions exprimées

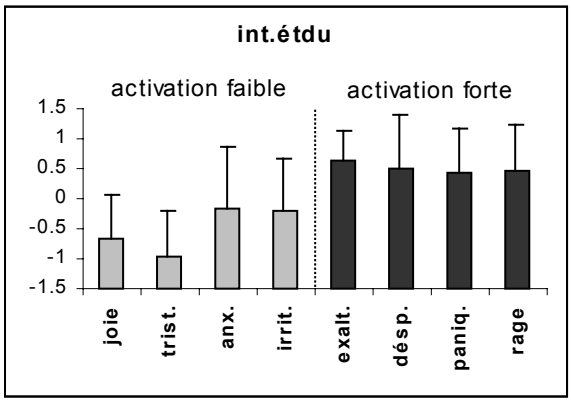
Les ANOVAs à mesures répétées présentées dans les tableaux 4 et 5 indiquent que, parmi les 10 paramètres acoustiques sélectionnés, 2 paramètres ne sont pas significativement affectés par les émotions exprimées. Il s'agit de la proportion d'énergie voisée comprise entre 300 et 500 Hz ($F(7, 56) = 1.09$, $p = .383$, $\eta^2 = .12$) et de la proportion d'énergie non-voisée en dessous de 1000 Hz ($F(7, 56) = 1.36$, $p = .241$, $\eta^2 = .15$). Les moyennes par émotion exprimée sont représentées pour les 8 autres paramètres dans les graphiques 1 à 8. Les moyennes pour les émotions comprenant un degré d'activation faible sont représentées par les barres claires, les barres foncées représentent les moyennes pour les émotions comprenant un degré d'activation fort. Les écarts-type sont représentés au-dessus des moyennes dans ces graphiques. Des ANOVAs ont été effectuées avec le facteur émotion exprimée comme unique prédicteur, les différences significatives selon les tests post hoc de Tukey (Tukey HSD, $p<.05$) sont rapportées sous les graphiques.

Graphique 1 : Moyennes par émotion exprimée pour l'intensité moyenne (standardisée par locuteur)



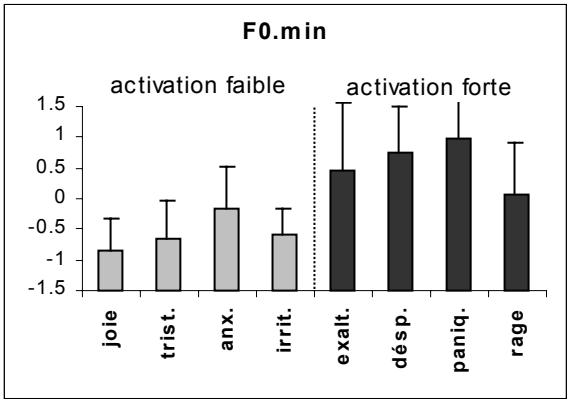
- trist. < anx., désp., paniq., irrit., exalt., rage
- joie, anx. < désp., paniq., irrit., exalt., rage
- désp., paniq., irrit. < rage

Graphique 2 : Moyennes par émotion exprimée pour l'étendue d'intensité (standardisée par locuteur)



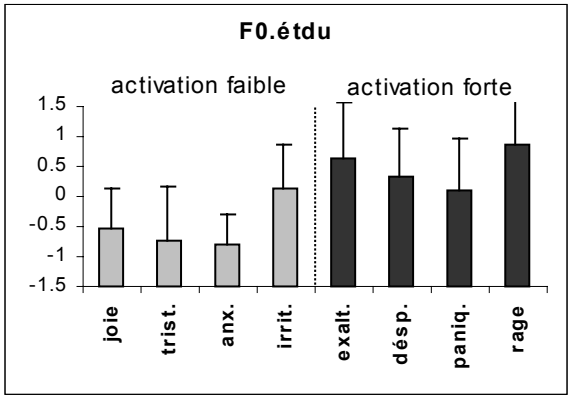
- joie, trist. < paniq., rage, désp., exalt.

Graphique 3 : Moyennes par émotion exprimée pour la F0 minimale (standardisée par locuteur)



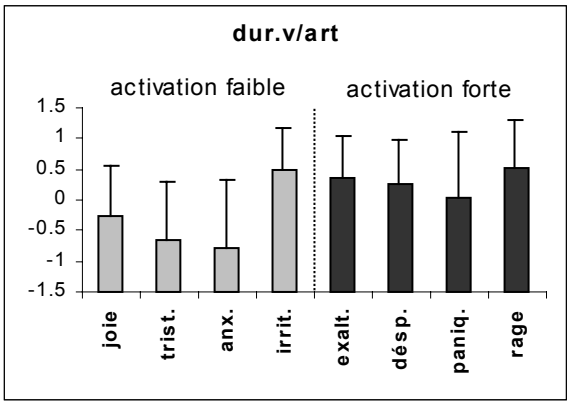
- joie < rage, exalt., désp., paniq.
- trist., irrit. < exalt., désp., paniq.
- anx. < désp., paniq.
- rage < paniq.

Graphique 4 : Moyennes par émotion exprimée pour l'étendue de F0 (standardisée par locuteur)



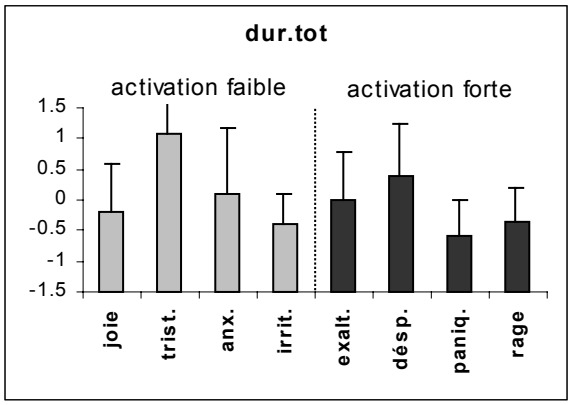
- anx., trist. < paniq., irrit., désp., exalt., rage
- joie < désp., exalt., rage.

Graphique 5 : Moyennes par émotion exprimée pour la durée proportionnelle des parties voisées sur les parties articulées (standardisée par locuteur)



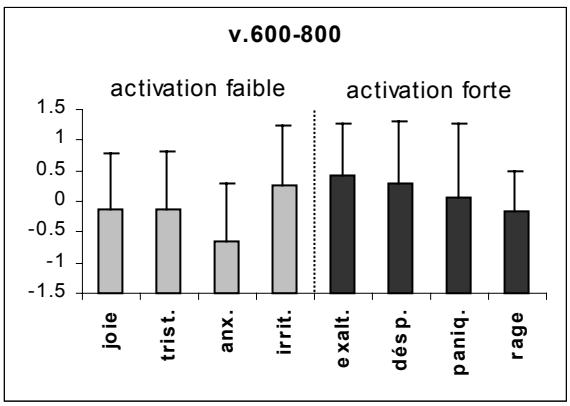
▪ anx., trist. < désp., exalt., irrit., rage

Graphique 6 : Moyennes par émotion exprimée pour la durée totale (standardisée par locuteur)



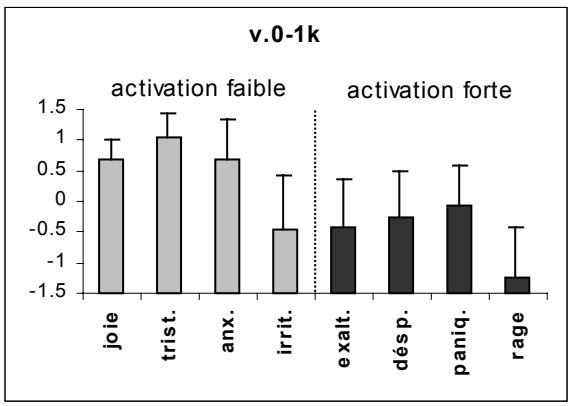
▪ paniq. < désp., trist.
▪ irrit., rage, joie, exalt., anx. < trist.

Graphique 7 : Moyennes par émotion exprimée pour la proportion d'énergie voisée de 600 à 800 Hz (standardisée par locuteur)



▪ anx. < exalt.

Graphique 8 : Moyennes par émotion exprimée pour la propor. d'énergie voisée en dessous de 1000 Hz (standardisée par locuteur)



▪ rage < irrit., exalt., désp., paniq., anx., joie, trist.
▪ irrit., exalt., désp., paniq. < anx., joie, trist.

Les graphiques 1 à 8 mettent en évidence des profils acoustiques différents pour chaque émotion exprimée. Toutefois, l'examen de ces graphiques révèle que les différentes émotions semblent être plus ou moins bien distinguées par ces mesures. La tristesse et la colère chaude sont notamment distinguées de la plupart des autres expressions sur plusieurs mesures. Les autres émotions semblent, à première vue, moins bien différenciées. Afin d'évaluer plus globalement la possibilité de discriminer les émotions exprimées à l'aide des paramètres acoustiques sélectionnés, une analyse discriminante a été effectuée. Les valeurs propres et le pourcentage de variance expliquée par les fonctions extraites par cette analyse sont présentées dans le tableau 9. La dernière colonne de ce tableau fourni une estimation de l'importance de la relation entre chaque fonction et les catégories (émotions exprimées) à discriminer. Selon le lambda de Wilks, seules les 3 premières fonctions discriminantes apportent une contribution significative à la discrimination.

Tableau 9 : Résultats de l'analyse discriminante: valeurs propres et % de variance expliquée par les fonctions discriminantes, relations entre les groupes et les fonctions (corr. canonique).

Fonction	valeur propre	% de variance	% cumulé de variance	corrélation canonique
1	3.641	73.0	73.0	.886
2	.820	16.5	89.5	.671
3	.340	6.8	96.3	.504
4	.118	2.4	98.7	.325
5	.042	.8	99.5	.200
6	.017	.3	99.9	.129
7	.006	.1	100.0	.078

Les 7 fonctions discriminantes dérivées des 8 paramètres acoustiques sélectionnés permettent de reclasser 60.4% des expressions dans leurs catégories d'origine. Lorsque les fonctions discriminantes sont calculées pour classer chaque expression en se basant sur l'ensemble des expressions excepté l'expression à classer ("cross validation"), 50% des expressions sont encore classées correctement. Le nombre d'expressions classées dans chaque catégorie par cette seconde analyse est représenté dans le tableau 10.

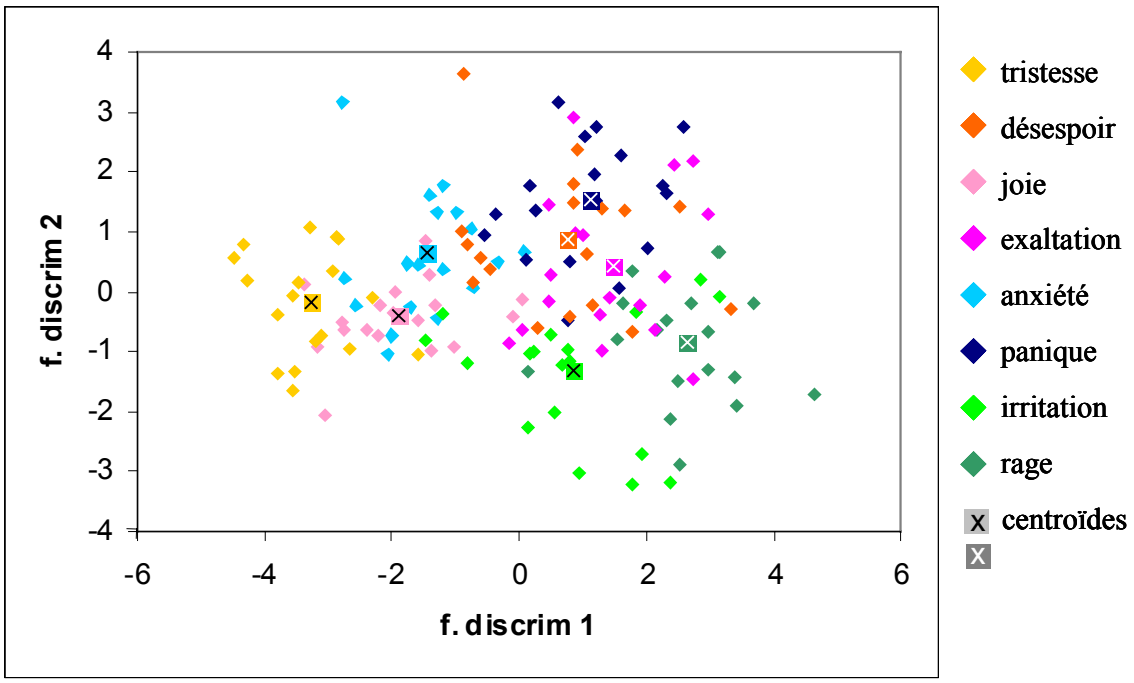
Tableau 10 : Classification des expressions par l'analyse discriminante (cross validated) basée sur les 8 paramètres acoustiques sélectionnés.

Groupe original	**Appartenance de groupe prédite**							
	1 (anx.)	2 (joie)	3 (rage)	4 (irrit.)	5 (paniq.)	6 (trist.)	7 (exalt.)	8 (désp.)
Anx.	**10**	5	0	0	0	2	0	1
Joie	2	**9**	0	2	0	5	0	0
Rage	0	0	**13**	1	0	0	3	1
Irrit.	1	1	4	**10**	0	0	2	0
Paniq.	4	0	0	1	**7**	0	5	1
Trist.	0	5	0	0	0	**13**	0	0
Exalt.	0	0	3	5	4	0	**3**	3
Désp.	3	0	1	1	1	0	5	**7**

Les expressions de tristesse (déprimée) et de colère chaude sont bien discriminées par les 8 paramètres acoustiques sélectionnés (13 expressions sur 18 sont regroupées correctement par l'analyse), alors que les expressions de joie exaltée ne sont manifestement pas aussi clairement caractérisées par ces mêmes paramètres acoustiques (les expressions de joie exaltée sont "confondues" avec les expressions de rage (colère chaude) et d'irritation (colère froide), ainsi qu'avec les expressions de peur panique et de désespoir). Le graphique 9 représente la position des 144 expressions, identifiées par le type d'émotion exprimée, sur les 2 premières fonctions discriminantes. Ce graphique montre le regroupement des expressions et les chevauchements des catégories d'origine qui résultent de l'analyse effectuée sur la base des 8 paramètres acoustiques sélectionnés. Pour simplifier et clarifier la présentation graphique des résultats, la contribution de la

3ème fonction à la discrimination n'est pas représentée, bien qu'elle soit statistiquement significative. Malgré cette réduction, la distribution des expressions sur les 2 premières fonctions reflète et clarifie les "confusions" rapportées dans le tableau 10. On observe que 3 des émotions faiblement activées – l'anxiété, la joie et la tristesse – sont relativement bien distinguées par l'analyse mais partiellement confondues entre elles. Les catégories panique, désespoir et exaltation sont relativement moins bien distinguées. Elles se "confondent" entre elles mais également avec l'anxiété (pour le désespoir et la panique) et la rage et l'irritation (pour l'exaltation). La rage et l'irritation sont également partiellement confondues avec les 2 autres émotions fortement activées mais malgré cela et malgré un "chevauchement" entre les 2 types de colère, elles sont mieux discriminées par l'analyse que la panique, le désespoir et l'exaltation.

Graphique 9 : Positions des expressions sur les 2 premières fonctions discriminantes et positions des centroïdes pour chaque émotion exprimée.



2.3.2.5 Rôle de l'activation émotionnelle

Les résultats décrits ci-dessus conduisent à s'interroger sur le rôle joué par l'activation émotionnelle dans la différenciation des émotions exprimées. La première fonction discriminante (v. graphique 9) semble distinguer en partie les émotions faiblement activées des émotions fortement activées. Les expressions qui comprennent un niveau d'activation faible tendent à avoir des scores plutôt faibles sur cette dimension alors que les expressions qui comprennent un niveau d'activation fort tendent à avoir des scores plutôt élevés sur cette dimension. Les expressions de colère froide (théoriquement peu activées) constituent une exception notable à cette observation, ils obtiennent en effet des scores relativement élevés sur cette dimension.

Afin d'estimer l'effet des émotions exprimées sur les 44 paramètres acoustiques mesurés indépendamment des 2 niveaux d'activation prédéfinis. Les ANOVAs à mesure répétées précédemment présentées (v. tableau 4 et 5) ont été re-calculées en ajoutant la variable 'activation' comme covariée. Le niveau "faible" (1) de cette variable correspond aux expressions appartenant aux catégories émotionnelles tristesse, anxiété, joie, irritation alors que le niveau "fort" (2) correspond aux expressions appartenant aux catégories émotionnelles désespoir, panique, exaltation, rage. Les résultats de ces ANOVAs sont présentés dans le tableau 11. Seul l'effet principal de l'émotion exprimée est représenté dans ce tableau. L'effet principal de la phrase et l'effet d'interaction (émotion x phrase) sont identiques aux résultats présentés dans les tableau 4 et 5.

Tableau 11 : ANOVAs à mesures répétées - Effets "within" de l'émotion sur les paramètres acoustiques standardisés par locuteur, avec covariation de la variable 'activation' (2 niveaux), les résultats significatifs (p<.05) sont en gras.

Code	source	df	F	sig.	eta^2	Code	source	df	F	sig.	eta^2
int.min	**emo**	**(7,56)**	**6.89**	**.000**	**0.46**	dur.s/tot	**emo**	**(7,56)**	**5.90**	**.000**	**0.42**
int.max	**emo**	**(7,56)**	**12.01**	**.000**	**0.60**	dur.n/tot	emo	(7,56)	0.95	.478	0.11
int.étdu	emo	(7,56)	1.48	.192	0.16	dur.v/tot	**emo**	**(7,56)**	**7.71**	**.000**	**0.49**
int.moy	**emo**	**(7,56)**	**15.05**	**.000**	**0.65**	dur.v/art	**emo**	**(7,56)**	**3.20**	**.006**	**0.29**
int.sd	emo	(7,56)	1.23	.300	0.13	dur.tot	**emo**	**(7,56)**	**6.12**	**.000**	**0.43**
F0.min	**emo**	**(7,56)**	**3.02**	**.009**	**0.27**	F0.05c	emo	(7,56)	1.61	.151	0.17
F0.max	emo	(7,56)	2.11	.058	0.21	F0.25c	emo	(7,56)	1.52	.179	0.16
F0.étdu	**emo**	**(7,56)**	**2.96**	**.010**	**0.27**	F0.50c	**emo**	**(7,56)**	**2.43**	**.030**	**0.23**
F0.moy	**emo**	**(7,56)**	**2.30**	**.040**	**0.22**	F0.75c	**emo**	**(7,56)**	**3.06**	**.008**	**0.28**
F0.sd	**emo**	**(7,56)**	**2.96**	**.010**	**0.27**	F0.95c	**emo**	**(7,56)**	**2.44**	**.030**	**0.23**
Hamm	**emo**	**(7,56)**	**5.59**	**.000**	**0.41**	F0.05-95	**emo**	**(7,56)**	**3.22**	**.006**	**0.29**
v.125-200	**emo**	**(7,56)**	**3.59**	**.003**	**0.31**	n.125-250	emo	(7,56)	0.73	.647	0.08
v.200-300	**emo**	**(7,56)**	**2.19**	**.049**	**0.21**	n.250-400	emo	(7,56)	1.08	.391	0.12
v.300-500	emo	(7,56)	1.05	.410	0.12	n.400-500	emo	(7,56)	0.58	.768	0.07
v.500-600	emo	(7,56)	1.45	.204	0.15	n.500-1k	**emo**	**(7,56)**	**2.50**	**.026**	**0.24**
v.600-800	emo	(7,56)	1.77	.111	0.18	n.1k-1.6k	emo[8]	(7,56)	2.44	.030	0.23
v.800-1k	**emo**	**(7,56)**	**2.38**	**.033**	**0.23**	n.1.6k-2.5k	emo	(7,56)	1.41	.221	0.15
v.1k-1.6k	**emo**	**(7,56)**	**9.20**	**.000**	**0.53**	n.2.5-4k	emo	(7,56)	1.70	.128	0.18
v.1.6k-5k	**emo**	**(7,56)**	**7.06**	**.000**	**0.47**	n.4k-5k	emo	(7,56)	0.82	.572	0.09
v.5k-8k	**emo**	**(7,56)**	**4.84**	**.000**	**0.38**	n.5k-8k	emo	(7,56)	0.39	.902	0.05
v.0-500	**emo**	**(7,56)**	**5.27**	**.000**	**0.40**	n.0-500	emo	(7,56)	1.94	.079	0.20
v.0-1k	**emo**	**(7,56)**	**9.64**	**.000**	**0.55**	n0-1k	emo	(7,56)	0.92	.498	0.10

Indépendamment de l'activation telle qu'elle a été prédéfinie (2 niveaux), la catégorie 'émotion' n'a pratiquement plus d'effet sur 2 des 8 variables qui avaient été sélectionnées pour l'analyse discriminante. L'émotion exprimée n'explique plus que 16% de la variance résiduelle de l'étendue de l'intensité des expressions et seulement 18% de la variance résiduelle de la proportion d'énergie contenue dans la bande fréquence allant de 600 à 800 Hz. Il reste toutefois des effets significatifs de

[8] Cet effet devient non significatif avec la correction de Greenhouse-Geisser: F(3.85, 30.83) = 2.44, p = .070

la catégorie 'émotion' sur 25 paramètres acoustiques. Les tailles de ces effets varient entre 21% de variance résiduelle expliquée par l'émotion pour la proportion d'énergie voisée comprise entre 200 et 300 Hz (soit des effets relativement faibles) et 65% de variance résiduelle expliquée par l'émotion pour l'intensité moyenne (soit un effet relativement important).

La même comparaison peut être appliquée aux fonctions de l'analyse discriminante. Une ANOVA effectuée pour évaluer l'effet de l'émotion exprimée (8 niveaux) sur la première fonction discriminante en contrôlant l'effet d'interaction entre le locuteur et l'émotion[9] révèle un effet significatif de l'émotion ($F(7, 56) = 48.97$, $p < .001$, $\eta^2 = .86$). Cet effet devient plus faible mais reste largement significatif lorsque l'activation (2 niveaux) est entrée comme covariée ($F(6, 56) = 22.04$, $p < .001$, $\eta^2 = .70$). L'émotion considérée sans contrôle du niveau d'activation a également un effet significatif sur la deuxième fonction discriminante ($F(7, 56) = 12.30$, $p < .001$, $\eta^2 = .61$) et sur la troisième fonction discriminante ($F(7, 56) = 4.91$, $p < .001$, $\eta^2 = .38$). Ces effets sont réduits lorsque le niveau d'activation est entré comme covariée. La taille de l'effet de l'émotion exprimée sur la deuxième fonction est alors de .55 ($F(6, 56) = 11.38$, $p < .001$, $\eta^2 = .55$) et la taille de l'effet de l'émotion sur la troisième fonction est de .36 ($F(6, 56) = 5.18$, $p < .001$, $\eta^2 = .36$). La réduction de l'effet de l'émotion par le contrôle du niveau d'activation est moins importante pour la deuxième et la troisième fonctions que pour la première fonction qui reflète donc effectivement, en partie mais non uniquement, l'effet de l'activation telle qu'elle a été définie théoriquement.

La question du rôle joué par l'activation dans la différentiation des émotions sur le plan acoustique étant particulièrement cruciale, quelques résultats plus détaillés relatifs à cette question sont présentés ci-dessous :

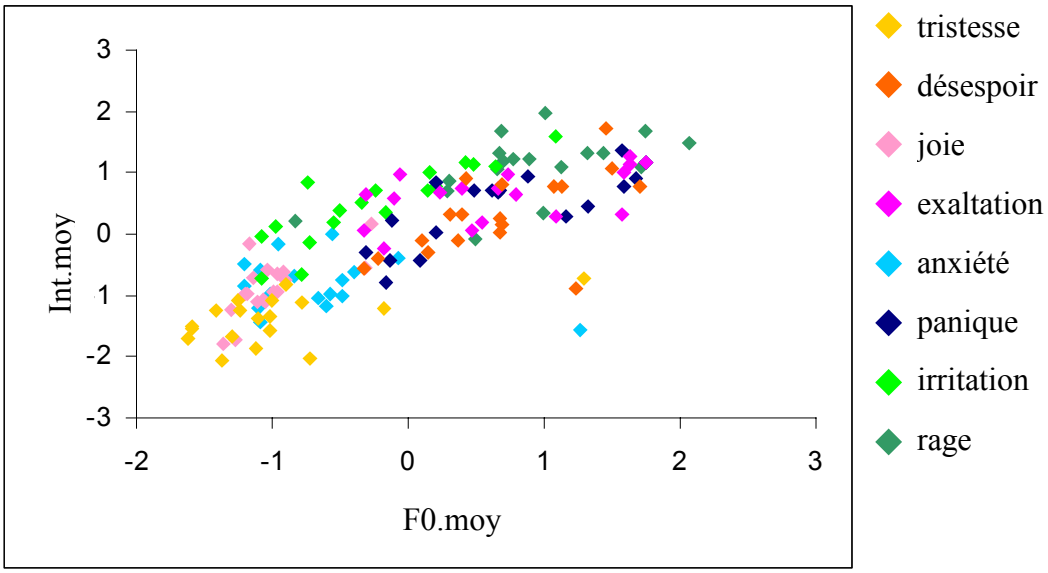
Certaines variables acoustiques reflètent en partie l'activation sous-jacente aux différentes émotions exprimées[10]. C'est le cas notamment de variables telles que la F0 moyenne ou l'intensité moyenne. L'échantillonnage d'expressions correspondant à des niveaux d'activation extrêmes a notamment pour conséquence d'accroître les corrélations entre ces variables acoustiques. Le graphique 10 illustre ce constat. Il représente la corrélation positive entre l'intensité moyenne et la F0 moyenne des expressions. On observe que la relation entre l'intensité moyenne et la F0 moyenne est le plus souvent moins importante lorsqu'on on la considère séparément pour chaque émotion exprimée –

[9] La variable 'locuteur' (identité de l'acteur) est entrée comme "random factor" dans l'analyse de variance.

[10] Bien entendu, cette affirmation n'exclut pas que ces variables soient affectées également par d'autres aspects émotionnels ou non.

les corrélations sont égales à -.27 pour la peur[11], .70 pour la joie, .57 pour la colère chaude, .87 pour la colère froide, .76 pour la peur panique, .51 pour la tristesse, .53 pour la joie intense, .61 pour le désespoir (ces corrélations sont calculées sur des échantillons de 18 expressions, la moyenne de ces 8 corrélations est égale à .54)[12] – que lorsqu'on la considère pour l'ensemble des expressions émotionnelles, dans ce cas la corrélation est égale à .79 (cette corrélation est calculée pour les 144 expressions).

Graphique 10 : Position des expressions – groupées en fonction de l'émotion exprimée – sur les paramètres acoustiques 'intensité moyenne' et 'F0 moyenne'.



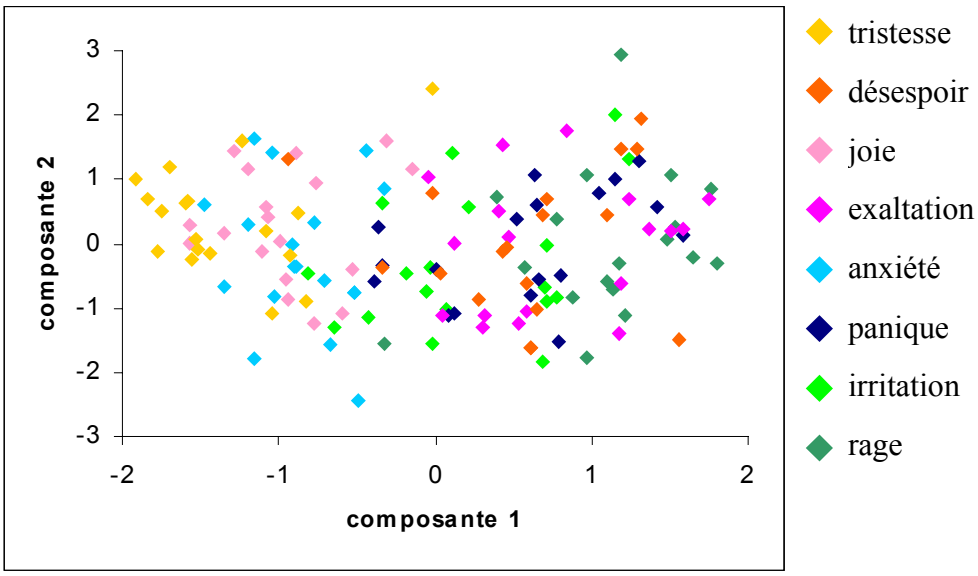
Les corrélations très élevées entre une partie des paramètres acoustiques mesurés (v. annexes A6.1 – A6.3) sont donc partiellement attribuables au fait que ces paramètres capturent le niveau d'activation sous-jacent aux expressions. Cette observation est corroborée par les résultats de l'analyse en composantes principales effectuée sur l'ensemble des paramètres acoustiques mesurés. Avant rotation, la première composante de l'ACP qui rend compte de 35.8 % de la variance des 44 paramètres acoustiques (v. tableau 6) reflète en partie l'activation sous-jacente aux expressions émotionnelles utilisées (v. graphique 11). Les expressions qui correspondent à la tristesse, à la joie et à l'anxiété reçoivent des scores plutôt faibles sur cette composante, alors que les expressions de désespoir, de panique, d'exaltation et de rage ont des scores factoriels élevés. Les expressions d'irritation font exception également dans cette analyse. Théoriquement peu activées, elles reçoivent des scores plus proches des émotions théoriquement fortement activées que des autres émotions théoriquement peu activées.

---

[11] La faible corrélation négative est due à un unique "outlier" et pourrait éventuellement correspondre à un problème d'identification de la F0 pour cette expression.

[12] Les annexes A7.1-A7.3 représentent les moyennes des corrélations entre les 44 paramètres acoustiques calculées séparément pour chaque émotion exprimée.

Graphique 11 : Position des expressions sur les 2 premières composantes de l'ACP avant rotation.



L'effet, évalué par une ANOVA, de l'émotion exprimée sur cette première composante en contrôlant l'effet d'interaction entre le locuteur et l'émotion est significatif ($F(7, 56) = 38.04$, $p < .001$, $\eta^2 = .83$) et diminue sensiblement lorsque le niveau d'activation (2 niveaux) est utilisé comme covariée ($F(6, 56) = 11.38$, $p < .001$, $\eta^2 = .55$). L'émotion exprimée n'a pas d'effet sur la deuxième composante de l'ACP représentée dans le graphique 11 ($F(7, 56) = 0.84$, $p = .557$, $\eta^2 = .10$).

2.4 Discussion/Conclusions

Avec 50% des expressions classées correctement par l'analyse discriminante ("cross validated"), la discrimination entre les catégories émotionnelles réalisée sur le plan acoustique est plus que satisfaisante. Ce résultat est en fait très proche de la discrimination observée dans les études de jugements effectuées avec la participation d'auditeurs humains. Les revues dans ce domaine rapportent un taux de reconnaissance correct des émotions exprimées avoisinant les 60%, tous types d'émotions confondus (e.g. Pittam & Scherer, 1993). La classification correcte de la moitié des expressions par l'analyse discriminante apparaît plus impressionnante encore si l'on considère qu'il s'agissait de différencier 8 catégories (sur la base de 18 occurrences par catégorie), la plupart des études réalisées avec des auditeurs humains ont en effet utilisé un nombre de catégories plus réduit, facilitant ainsi la discrimination. En terme de 'profils acoustiques', on notera que le type d'émotion conserve un effet sur la majorité des paramètres acoustiques après que le niveau d'activation ait été contrôlé, il semble donc qu'il serait possible d'identifier des profils différents pour différents types d'émotions ayant le même niveau d'activation à condition d'inclure un nombre suffisant de paramètres acoustiques relativement indépendants (v. graphiques 1 à 8, des différences significatives apparaissent entre les émotions de même activation).

D'un autre côté, l'analyse discriminante "confond" plus souvent les émotions de même niveau d'activation relativement aux émotions incluant différents niveaux d'activation qui sont mieux discriminées les unes des autres. Sur le plan des profils acoustiques, on observe également pour la grande majorité des paramètres acoustique un effet très important du niveau d'activation. Cet effet, dans beaucoup de cas – notamment pour la plupart des mesures de F0 – est en réalité supérieur à l'effet du type d'émotion. Les résultats présentés ci-dessus indiquent donc que les paramètres acoustiques mesurés capturent effectivement une dimension sous-jacente aux émotions. Cette dimension est déjà apparente dans les corrélations, souvent plus importantes que la normale, entre les différents paramètres acoustiques. L'examen de la 'réduction' des corrélations sur la 1ère composante de l'ACP nous autorise à interpréter cette dimension comment correspondant, au moins en partie, à l'activation sous-jacente aux émotions.

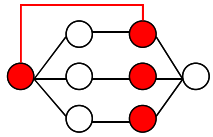
Toutefois, la 1ère composante de l'ACP (tout comme la première fonction discriminante) permet encore de différencier partiellement les catégories émotionnelles une fois que l'effet de l'activation est contrôlé. De plus, une partie des autres composantes de l'ACP (ainsi que la 2ème et la 3ème fonction de l'analyse discriminante) qui rendent compte d'une part importante de la variance des paramètres acoustiques sont d'avantage affectés par le type d'émotion que par le niveau d'activation. Selon ces dernières observations, les paramètres acoustiques permettraient donc de différencier les émotions au-delà de leur niveau d'activation. Cette conclusion est malheureusement limitée par la manière dont l'activation a été opérationnalisée. En effet, **si l'on accepte l'hypothèse** que les variables acoustiques mesurent essentiellement l'activation émotionnelle, il apparaît que la définition de l'activation utilisée (2 niveaux, faible versus forte) n'est probablement pas suffisamment différenciée. Si l'on décide d'interpréter la première fonction discriminante comme une dimension sous-jacente qui reflèterait l'activation, la tristesse tendrait à être moins activée que la joie et la peur qui elles-mêmes tendraient à être moins activées que la joie intense, la peur panique, le désespoir et la colère froide qui seraient moins activées que la colère chaude. Cette argumentation peut être poussée jusqu'à postuler que les catégories émotionnelles utilisées (8 émotions) ne définissent pas des niveaux d'activation suffisamment différenciés, des réalisations spécifiques de chaque type d'émotion – par exemple de la 'colère froide' ou de la 'peur anxieuse' – pourraient en effet correspondre à des états plus ou moins activés.

Il apparaît que le rôle joué par l'activation émotionnelle ne peut être entièrement défini dans le cadre de l'approche adoptée dans cette étude. Cette problématique pourrait être abordée avec plus de succès en utilisant une définition dimensionnelle des réactions émotionnelles. Des expressions correspondant non pas à des catégories/familles émotionnelles mais à des dimensions telles que l'activation, la valence ou le contrôle (v. (Schlossberg, 1954)) devraient être enregistrées, en prenant

soin d'inclure pour chaque dimension (notamment pour la dimension d'activation) plusieurs niveaux différents. La définition explicite du niveau d'activation lors de l'encodage permettrait d'établir l'influence de cette dimension (relativement à d'autres dimensions émotionnelles) sur les caractéristiques acoustiques de la voix. La manipulation indépendante de plusieurs dimensions émotionnelles et l'enregistrement d'expressions vocales qui correspondraient aux états résultant de cette manipulation semblent toutefois difficilement réalisables. Une autre alternative pourrait consister à mesurer un indicateur indépendant du degré d'activation émotionnel lors de l'enregistrement des expressions vocales. Une mesure du degré d'activation physiologique par exemple – telle que la mesure du rythme cardiaque – qui pourrait être utilisée afin d'évaluer l'influence de l'activation sur le plan acoustique (v. Johnstone, 2001).

En conclusion, les résultats présentés ci-dessus indiquent que – dans la limite de la définition relativement grossière qui a été adoptée pour le niveau d'activation – il est possible de différencier sur le plan acoustique non seulement le niveau d'activation émotionnel mais également le type d'émotion exprimé. La possibilité qu'un degré d'activation plus finement différencié soit exprimé par les acteurs et capturé par les mesures acoustiques ne peut être exclue et devrait faire l'objet de plus d'intérêt dans les études futures.

3 Caractéristiques vocales perçues des émotions exprimées



3.1 Introduction

Dès 1978, Scherer a proposé d'utiliser le modèle 'lens' de Brunswik (Scherer, 1978, 2003; Brunswik, 1956) comme paradigme pour la recherche sur la communication non-verbale des émotions (v. introduction théorique, section 1.3). Le modèle 'lens' adapté par Scherer (1978) distingue quatre étapes dans le processus de communication: premièrement l'état interne (émotion) du locuteur; deuxièmement l'extériorisation de cet état au travers des caractéristiques acoustiques de la voix qui encodent l'émotion du locuteur; troisièmement la perception de ces caractéristiques vocales par un auditeur; quatrièmement l'inférence par l'auditeur de l'état émotionnel du locuteur à partir des caractéristiques vocales perçues. Dans le chapitre qui suit, une méthode destinée à évaluer la troisième étape de ce processus est présentée. Des jugements relatifs à différentes caractéristiques vocales perçues ont été obtenus à l'aide de cette méthode. Les caractéristiques vocales ainsi évaluées sont mises en relation avec les émotions exprimées et avec différentes caractéristiques acoustiques des expressions utilisées. Des aspects de la validité et de la pertinence de cette approche sont discutés.

3.1.1 Evaluer les caractéristiques vocales perçues - pourquoi

Un problème central dans les travaux sur la communication émotionnelle vocale concerne le manque de connaissances clairement spécifiées relativement aux liens entre les caractéristiques objectives du signal acoustique et la perception que peuvent en avoir les auditeurs.

On sait que l'évaluation auditive de certaines caractéristiques vocales, telles que l'intensité ou la hauteur, ne corrèle pas parfaitement avec une mesure objective simple de ces mêmes caractéristiques, par exemple l'intensité acoustique ou la fréquence fondamentale. Ceci est dû partiellement au fait que les impressions auditives sont influencées par diverses attentes et habitudes formées culturellement dans un contexte linguistique donné, mais surtout au fait que le système auditif ne fonctionne pas comme un instrument d'enregistrement et de mesure. Le système auditif humain est peu performant lorsqu'il s'agit d'isoler des paramètres acoustiques individuels. Les jugements concernant par exemple la hauteur de la voix sont affectés non seulement par la fréquence fondamentale (F0), mais aussi par d'autres facteurs tels que la distribution de l'énergie dans le spectre (Scherer, 1982). Dans le même sens, l'intensité perçue (‘loudness’) correspondrait d'avantage à l'effort vocal qu'à l'intensité acoustique (dérivée de l'amplitude du signal acoustique). Selon Frick (1985), l'intensité ne serait un bon indicateur de l'effort vocal que lorsqu'elle est

mesurée au point de l'émetteur, il est donc fort probable que les auditeurs utilisent d'autres indicateurs pour évaluer l'effort vocal.

Si les propriétés perçues de la voix ne correspondent pas directement à des propriétés physiques isolées du signal lorsqu'il s'agit de caractéristiques relativement clairement définies sur le plan acoustique (telle que la hauteur/F0 ou l'intensité), la description physique d'autres propriétés perçues de la voix est encore plus mal définie. Dans une revue de la recherche sur la qualité de la voix pathologique, Kreiman souligne que la perception de la voix ne peut être réduite directement aux propriétés du signal (Kreiman, 1998). Les résultats des recherches qu'elle mentionne indiquent que différents auditeurs utilisent différentes stratégies perceptives et, d'autre part, que les stratégies varient en fonction des propriétés du signal. L'attention accordée à des aspects particuliers d'un signal – par exemple la qualité rauque (‘roughness’) ou soufflée (‘breathiness’) étudiées par Kreiman – serait notamment partiellement prédite par le degré de variabilité de ces aspects. L'attention d'un auditeur, par exemple pour la qualité soufflée d'une voix, serait particulièrement orientée vers cet aspect dans un contexte où des échantillons vocaux présentent des modifications importantes au niveau de la qualité soufflée.

Dans le même sens, on peut émettre l'hypothèse que les caractéristiques perçues de la voix émotionnelle ne dépendent pas uniquement et linéairement des caractéristiques objectives des expressions produites par les locuteurs. La qualité vocale perçue pour une même expression varie probablement en fonction des auditeurs et de leurs stratégies perceptives; et la saillance perceptive de différents aspects des expressions pourrait être conditionnée par d'autres aspects de ces expressions.

Suivant les considérations ci-dessus, si l'on veut aboutir à une meilleure compréhension du processus d'attribution émotionnelle à partir de l'expression vocale et du rôle joué par différentes caractéristiques vocales dans ce processus, il paraît important d'obtenir des informations sur la manière dont l'expression vocale est perçue et catégorisée par les auditeurs. L'objectif de cette deuxième partie de la thèse sera donc d'ajouter à l'étude des caractéristiques objectives des expressions émotionnelles (caractéristiques acoustiques examinées dans la première partie) une étude des caractéristiques perçues des expressions émotionnelles.

3.1.2 Evaluer les caractéristiques vocales perçues - comment

Très peu d'études ayant jusqu'ici cherché à évaluer la qualité vocale perçue de la voix émotionnelle, il n'existe pas d'instrument validé et généralement accepté à cet effet. Dans le cadre de la recherche sur la voix pathologique, la qualité vocale est couramment évaluée à l'aide de questionnaires comprenant un nombre variable d'échelles représentant des dimensions de la qualité vocale.

Kreiman & Gerratt (1998) rapportent les problèmes liés à l'utilisation de ces dimensions (échelles) identifiées par des labels verbaux pour mesurer la qualité de la voix. En particulier, ils soulignent que lorsque l'on mesure la qualité vocale en enregistrant des évaluations sur des échelles qui représentent des aspects particuliers de la qualité, on implique que l'impression totale qu'un auditeur forme lorsqu'il entend une voix peut être décomposée en plusieurs aspects distincts qui correspondraient à différents termes tels que ‘rauque’, ‘grinçant’, ‘soufflé’. Kreiman et Gerrat mettent en doute la compétence des locuteurs à isoler de telles dimensions et à les évaluer séparément. Kreiman (1998) met également en évidence le fait que les dimensions habituellement soumises à l'évaluation (‘breathiness’, ‘roughness’ etc...) ne sont pas le produit d'une théorie cohérente de la perception de la voix. Souvent les dimensions sont choisies/définies en dehors de tout support théorique ; dans les cas plus favorables où un modèle sous-jacent aux dimensions utilisées existe, il s'agit généralement d'un modèle de la production vocale. Selon Kreiman, il n'existe pas de modèle de la perception de la voix et donc aucun support théorique au choix des dimensions à inclure pour l'évaluation de la qualité vocale globale.

Le même auteur (Kreiman, Gerratt, Precoda, & Berke, 1992) a par ailleurs montré, pour une tâche d'évaluation de la qualité de voix pathologiques, dans laquelle les auditeurs devaient effectuer des jugements de similarité, que la similarité perçue n'était pas constante entre les auditeurs (absence d'accord inter-juges). Dans une autre étude, Kreiman (1998) a évalué l'accord inter-juges concernant l'évaluation de voix sur des échelles représentant des dimensions de la qualité vocale ("roughness" et "breathiness"). La probabilité d'accord sur ces dimensions n'était élevée que pour les jugements extrêmes (c.à.d. pour les jugements moyens très bas ou très élevés sur les échelles considérées), dans cette étude des scores moyens situés autour du milieu des dimensions évaluées ne représentent pas des indicateurs d'une qualité moyenne perçue concernant ces dimensions mais une absence d'accord inter-juges.

En conséquence, une partie importante de l’étude présentée ci-dessous a consisté à développer une procédure pour l'évaluation des caractéristiques perçues de la voix et de la parole émotionnelle en tenant compte des objections issues de la recherche sur l'évaluation de la voix pathologique. Le choix des dimensions (échelles) ne pouvant être motivé par une théorie de perception de la voix, c'est un modèle de production qui a servi de base à ce travail. Sangsue et al., (1997) ont initié une démarche dans ce sens en développant un questionnaire pour l'évaluation des caractéristiques vocales perçues à partir du modèle de production vocale proposé par Laver (1980). Ce questionnaire a été repris et modifié en tenant compte : (a) des résultats issus des premières applications du questionnaire à des échantillons de voix émotionnelle et (b) des résultats obtenus dans une série de tests effectués afin de vérifier la compréhension des termes choisis pour désigner

les dimensions retenues et dans le but d'identifier éventuellement d'autres dimensions qui pourraient être ajoutées à ce questionnaire. Afin de pallier aux objections relatives à l'absence d'accord inter-juges, une procédure ad hoc pour recueillir les jugements a ensuite été développée sur la base d'une méthode proposée par Granqvist (1996). Cette procédure a été prétestée pour une partie des dimensions (échelles) sélectionnées et a été ensuite utilisée pour évaluer les caractéristiques vocales perçues des expressions émotionnelles analysées dans la première partie de la thèse. Les procédures utilisées ainsi que les résultats obtenus sont décrits ci-dessous.

3.2 Méthode

La principale difficulté lorsque l'on souhaite obtenir des jugements relatifs aux caractéristiques vocales perçues réside dans l'absence de termes qualificatifs de la voix et de la parole dans le vocabulaire d'usage courant. Une série de questionnaires et de tests utilisés afin d'évaluer le vocabulaire spontané relatif à la voix, ainsi que la compréhension de différents termes utilisés par des professionnels de la voix – en particulier les termes proposés dans le questionnaire de Sangsue et al. (1997) sur la base du modèle de production de Laver et, également, des termes traduits de Laver (1980) – nous ont permis d'établir que les auditeurs inexperts ne disposent que d'un vocabulaire limité et ne distinguent explicitement (verbalement) qu'un très petit nombre de dimensions vocales. La description et les résultats de ces prétests figurent en annexe B1. A l'issue de ces prétests, neuf dimensions vocales ont été sélectionnées pour être utilisées dans les études de jugements relatives aux caractéristiques vocales de la voix émotionnelle.

Une seconde difficulté a été mise en évidence par un groupe de chercheurs dans le domaine de l'étude de la voix pathologique. Comme mentionné ci-dessus, Kreiman & Gerratt (1998) ont démontré que l'évaluation de la qualité vocale sur des échelles telles que "rauque" ("rough") ou "soufflée" ("breathy") ne produit pas de jugements fiables. Kreiman & Gerratt contestent à la fois la fidélité test re-test et la fidélité interindividuelle de ce type de jugements. Selon ces auteurs les standards de comparaison utilisés par les auditeurs lorsqu'ils effectuent des jugements varient d'un auditeur à l'autre et varient également dans le temps pour un même auditeur.

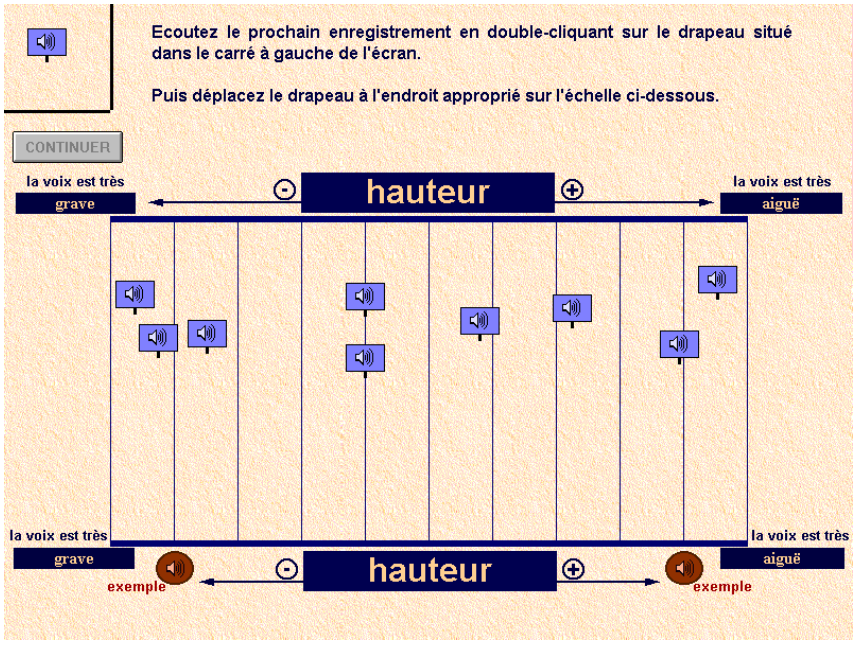
Afin de palier d'une part au problème de la définition des termes utilisés pour qualifier la voix et d'autre part au problème des standards de comparaison instables, nous avons adopté et modifié une approche proposée par Granqvist (1996).

3.2.1 Procédure développée pour les jugements des caractéristiques vocales perçues

Dans une approche conventionnelle, les enregistrements devant être jugés sont présentés dans un ordre aléatoire et sont évalués par des auditeurs sur un ensemble d'échelles (dimensions vocales)

immédiatement après la présentation de chaque enregistrement. Dans l'approche que nous avons adoptée, une échelle est présentée à l'auditeur sur un écran d'ordinateur (figure 1). La tâche de l'auditeur consiste à disposer les enregistrements sur cette échelle. Tous les enregistrements produits par un locuteur donné apparaissent sur l'écran dans un ordre aléatoire sous forme d'icônes identiques, l'auditeur les dispose sur l'échelle en fonction de la valeur qu'il souhaite leur attribuer sur cette échelle. Il est libre de ré-écouter les enregistrements aussi souvent qu'il le souhaite et peut modifier ses réponses lorsqu'il le juge nécessaire. Les enregistrements produits par différents locuteurs sont présentés sur des échelles séparées afin que les jugements réalisés ne soient pas conditionnés par les différences vocales entre les locuteurs mais uniquement par les différences liées aux émotions exprimées par les locuteurs[13]. Les échelles sont présentées séquentiellement, dans un ordre aléatoire différent pour chaque auditeur. Les réponses sont enregistrées par l'ordinateur sur une échelle continue allant de 0 (pour un enregistrement positionné sur la barre à l'extrême gauche de l'échelle) à 10 pour un enregistrement positionné sur la barre à l'extrême droite de l'échelle).

Figure 1: Illustration de la méthode utilisée. La dimension vocale 'hauteur' est représentée par une échelle bipolaire (la voix est très grave ↔ très aiguë).



La possibilité de comparer directement un ensemble d'expressions pour une dimension donnée et de les évaluer relativement aux autres expressions répond au problème de la modification dans le temps des standards de comparaison internes des auditeurs. Afin d'adresser le problème du partage

[13] La présentation séparée des expressions produites par chaque locuteur empêche la comparaison directe des expressions produites par différents locuteurs. Elle permet d'éviter par exemple que la voix (féminine) d'une locutrice soit systématiquement évaluée comme plus aiguë que celle d'un locuteur (masculin). En revanche l'évaluation des variations de hauteur produite par un même locuteur/trice d'une expression à l'autre sera favorisée par cette procédure.

de la définition et du standard de comparaison entre différents auditeurs, deux exemples extrêmes ont été enregistrés pour chaque dimension. Les deux enregistrements sont présentés en même temps que l'échelle qu'ils illustrent et permettent d'uniformiser la compréhension de la dimension représentée par l'échelle pour différents auditeurs. Ces enregistrements peuvent également être ré-écoutés à tout moment durant la procédure de jugement, l'auditeur est toutefois instruit de ne les utiliser qu'à fin d'exemple et non à fin de comparaison directe avec les enregistrements qu'il doit positionner sur l'échelle. Un prétest de cette procédure a été réalisé pour une partie des 9 dimensions vocale sélectionnées.

3.2.2 Prétest de la procédure de jugement

Pour ce prétest, un ensemble d'expressions émotionnelles produites par deux locuteurs ont été sélectionnées parmi les expressions produites par les acteurs de l'étude dirigée par Klaus Scherer, Harald Wallbott, Rainer Banse et Heiner Ellgring (une description de ces expressions est présentée dans la section 2.2.1). Les deux acteurs ont été choisis sur la base des indications publiées par Banse & Scherer (1996) relativement à la qualité des expressions émotionnelles produites par les acteurs. Huit états émotionnels ont été sélectionnés parmi les 14 états simulés par les acteurs. La sélection a été effectuée aléatoirement avec la restriction suivante: des états à la fois positifs et négatifs, faiblement et fortement activés, ont été inclus afin de favoriser la variabilité des expressions. Deux expressions ont été ensuite sélectionnées pour chacun des deux acteurs et pour chacun des huit états émotionnels, des expressions présentant une bonne qualité acoustique ont été choisies. La sélection des expressions utilisées est représentée dans le tableau 1. Le locuteur qui a produit les exemples destinés à uniformiser la compréhension et les standards de comparaison interindividuels ne fait pas partie des locuteurs dont les expressions sont évaluées. Il prononce à chaque fois la phrase « je ne peux pas le croire ».

Lors du prétest, deux expressions (une expression de joie intense, U11103, produite par le 1er acteur et une expression de peur panique, P12212, produite par le 2ème acteur) ont été choisies aléatoirement parmi les expressions sélectionnées. Ces deux expressions ont été présentées à 2 reprises pour chacune des dimensions jugées. Ces répétitions ont été effectuées afin d'évaluer la consistance des réponses des participants. Toutes les expressions (y compris les 2 expressions répétées) ont été présentées dans un ordre aléatoire redéfini par l'ordinateur pour chaque participant. Toutefois, après la présentation de la dernière expression, les 17 expressions (16 expressions dont une est répétée) produites par chaque acteur restent présentes simultanément sur l'écran/échelle pour chaque dimension jugée.

Tableau 1 : Expressions sélectionnées pour le prétest de la procédure de jugement des caractéristiques vocales

	dégoût	dégoût	colère chaude	colère chaude	intérêt	intérêt	colère froide	colère froide
acteur3	E11203	E22103	H11203	H12203	I11103	I12103	K11103	K22103
acteur12	E12112	E12212	H12112	H21212	I11212	I12212	K11212	K21212
	ennui	ennui	peur panique	peur panique	tristesse	tristesse	joie intense	joie intense
acteur3	L21203	L22203	P11103	P12103	T21203	T22103	U11103	U22203
acteur12	L11212	L12212	P12212	P21112	T11112	T12112	U11212	U12112

La 1ère lettre et le 1er chiffre correspondent au scénario utilisé pour définir l'émotion exprimée (v. annexe A2), le 2ème chiffre correspond à la "phrase" prononcée, le 3ème chiffre correspond à la répétition de l'enregistrement et les 2 derniers chiffres au code de l'acteur (v. section 2.2.1 pour plus de détails).

Dans la procédure utilisée, les expressions peuvent être ré-écoutées et les réponses modifiées de nombreuses fois. Elle est en conséquence plus coûteuse en temps qu'une procédure conventionnelle. Les différences interindividuelles sur le plan de la latence des réponses sont également plus importantes avec cette procédure qu'avec une procédure classique dans laquelle les temps de réponse sont en général limités par un rythme prédéfini de présentation des expressions. A raison de deux locuteurs prononçant 17 expressions chacun, trois dimensions vocales (soit 102 décisions et positionnements des expressions sur les échelles) peuvent toutefois être évaluées dans un laps de temps relativement court.

Trois groupes de participants ont été recrutés successivement. Le premier groupe, composé de 9 étudiants de l'Université de Genève (4 hommes et 5 femmes, âge moyen = 24 ans, écart-type = 5.8 ans) a évalué les enregistrements décrits ci-dessus sur les 3 dimensions vocales suivante[14]: 'intensité' (*volume*, la voix est très *faible* à très *forte*), 'hauteur' (*hauteur*, la voix est très *grave* à très *aiguë*) et 'intonation' (*mélodie*, la voix est très *monotone* à très *modulée*). Le deuxième groupe composé de 9 autres étudiants de l'université de Genève (3 hommes et 6 femmes, âge moyen = 25 ans, écart-type = 9.8 ans) a évalué les enregistrements décrits ci-dessus sur les 3 dimensions vocales suivante : 'articulation' (*articulation*, la parole est très *bien articulée* à très *mal articulée*), 'fluidité' (*fluidité*, la parole est très *hachée* à très *fluide* et 'rapidité' (*vitesse*, la parole est très *lente* à très *rapide*). Finalement, un dernier groupe d'étudiants (4 hommes et 12 femmes, âge moyen = 25 ans, écart-type = 7.3 ans) a été recruté afin d'évaluer la possibilité d'utiliser cette procédure pour obtenir des jugements relatifs à la qualité émotionnelle des expressions. Ce groupe a évalué les expressions pour l'intensité de 'colère' (*aucune colère* à *colère extrême*), l'intensité de 'joie' (*aucune joie* à *joie extrême*) et l'intensité de 'tristesse' (*aucune tristesse* à *tristesse extrême*).

[14] Les parenthèse indiquent les termes qui ont été utilisés pour qualifier les échelles présentées aux participants.

Pour chaque groupe, les expressions produites par l'un des deux locuteurs sont sélectionnées aléatoirement et un ordre aléatoire de présentation des 3 dimensions (vocales ou émotionnelles) est défini par l'ordinateur pour l'évaluation des expressions du premier locuteur sélectionné. Un second ordre aléatoire de présentation des 3 dimensions (vocales ou émotionnelles) est ensuite défini pour l'évaluation des expressions produite par l'autre locuteur.

3.2.2.1 Résultats du prétest

Fidélité et consistance des jugements

La fidélité inter-juge est globalement très élevée. Le tableau 2 représente les coefficients de fidélité (intraclass correlations, ICC) pour toutes les dimensions (vocales et émotionnelles) évaluées. La première valeur (r) correspond à l'évaluation de la corrélation moyenne entre les juges, la deuxième valeur (R) correspond au coefficient de fidélité ajusté relativement au nombre de corrélations (i.e. au nombre de juges). Toutes les corrélations moyennes sont significatives ($p<.001$). Les coefficients de fidélité sont égaux ou supérieurs à .8 pour toutes les dimensions évaluées. En conséquence, les jugements moyens peuvent être considérés comme fiables et peuvent être utilisés afin de caractériser les expressions jugées. Certains jugements apparaissent toutefois comme moins fiables que d'autres. En particulier, les jugements d'*articulation* et de *fluidité* qui présentent des corrélations inter-juges moyennes relativement plus faibles que les autres jugements.

Tableau 2 : Coefficients de fidélité (r = single measure intraclass correlation, R = average measure intraclass correlation) pour les dimensions vocales et émotionnelles évaluées.

	Groupe 1 (N = 9)			Groupe 2 (N = 9)			Groupe 3 (N = 16)		
	hauteur	intensité	intonat.	articul.	fluidité	rapidité	colère	joie	tristesse
r	.57	.89	.52	.39	.31	.77	.69	.55	.63
R	.92	.99	.91	.85	.80	.97	.97	.95	.96

La consistance des jugements a été évaluée en calculant, pour chaque juge et pour chaque échelle, la distance moyenne entre les jugements correspondant aux deux expressions dupliquées[15] et la distance moyenne entre toutes les autres expressions. Pour chacune de ces deux valeurs, une moyenne et une médiane ont été ensuite calculées pour chaque groupe de juges. Une faible minorité de juges présentant des résultats atypiques (outliers), les médianes constituent de meilleurs résumés des résultats au niveau des groupes. Les moyennes pour chaque juge, ainsi que les moyennes et les médianes pour chaque groupe de juges sont représentées dans le tableau 3.

[15] Moyenne de la valeur absolue de la différence entre le jugement pour l'expression U11103 et U11103' et de la valeur absolue de la différence entre le jugement pour l'expression P12212 et P12212'.

On observe que la distance moyenne entre les expressions dupliquées est nettement plus faible que la distance moyenne entre les autres expressions pour la majorité des juges et pour l'ensemble des dimensions. Pourtant, quelques différences apparaissent. Certains juges placent les expressions identiques à une distance proche (voir supérieure) à la distance moyenne à laquelle ils placent les autres expressions. Il est possible que les juges pour lesquels cette observation se répète pour les 3 dimensions jugées (par exemple v6 et v11) répondent au hasard. La comparaison des médianes au niveau des groupes de juges révèle toutefois une bonne consistance sur le plan général pour la plupart des échelles. À l'exception peut-être de la dimension *fluidité* qui semble être évaluée de manière moins consistante que les autres dimensions. Il est intéressant de relever que les jugements émotionnels ne sont pas plus fidèles (entre les juges) ni plus constants (pour les mêmes juges) que la plupart des jugements de qualité vocale.

Tableau 3 : Consistance des jugements : comparaison de la distance moyenne* entre les expressions dupliquées avec la distance moyenne* entre toutes les autres expressions.

INTENSITÉ	v1	v2	v3	v4	v5	v6	v7	v8	v9									moy.	méd.
Δ moy. autres	3.4	3.4	3.8	4.0	3.5	3.4	3.7	3.9	3.6									3.6	3.6
Δ moy. dupliqués	0.1	0.6	0.0	0.0	0.0	3.7	0.2	0.0	0.0									0.5	0.0
HAUTEUR	v1	v2	v3	v4	v5	v6	v7	v8	v9									moy.	méd.
Δ moy. autres	3.1	3.5	3.2	3.6	3.5	3.6	3.4	3.4	3.5									3.4	3.5
Δ moy. dupliqués	0.6	0.0	0.4	0.5	0.0	2.2	1.2	0.0	0.0									0.5	0.4
INTONATION	v1	v2	v3	v4	v5	v6	v7	v8	v9									moy.	méd.
Δ moy. autres	3.2	3.3	3.4	3.7	3.6	3.5	3.6	2.7	3.4									3.4	3.4
Δ moy. dupliqués	1.5	0.0	0.0	0.1	0.5	2.9	1.6	0.1	0.0									0.7	0.1
ARTICUL.	v10	v11	v12	v13	v14	v15	v16	v17	v18									moy.	méd.
Δ moy. autres	3.1	4.4	3.0	3.2	3.1	2.2	4.1	3.3	3.8									3.4	3.2
Δ moy. dupliqués	0.0	2.4	0.8	0.4	0.5	0.2	1.0	1.0	0.9									0.8	0.8
RAPIDITÉ	v10	v11	v12	v13	v14	v15	v16	v17	v18									moy.	méd.
Δ moy. autres	3.6	3.6	2.7	2.4	3.2	2.5	3.5	3.2	3.1									3.1	3.2
Δ moy. dupliqués	0.9	4.1	0.4	0.5	2.5	0.0	1.5	0.0	1.5									1.3	0.9
FLUIDITÉ	v10	v11	v12	v13	v14	v15	v16	v17	v18									moy.	méd.
Δ moy. autres	3.5	3.6	2.8	2.9	3.5	2.1	3.7	3.0	3.3									3.2	3.3
Δ moy. dupliqués	1.3	1.4	1.4	0.4	4.0	0.5	1.4	1.3	3.3									1.7	1.4
COLÈRE	e1	e2	e3	e4	e5	e6	e7	e8	e9	e10	e11	e12	e13	e14	e15	e16		moy.	méd.
Δ moy. autres	3.8	3.1	3.0	3.7	3.4	2.8	3.6	3.8	3.5	3.6	3.8	4.1	3.8	3.2	3.8	2.6		3.5	3.6
Δ moy. dupliqués	4.0	0.0	0.1	0.1	0.1	0.1	0.1	0.6	0.9	0.6	1.4	0.0	0.1	0.2	1.2	0.0		0.6	0.1
JOIE	e1	e2	e3	e4	e5	e6	e7	e8	e9	e10	e11	e12	e13	e14	e15	e16		moy.	méd.
Δ moy. autres	3.7	2.4	2.3	2.6	1.8	2.2	1.7	1.2	1.7	3.2	2.0	2.9	3.5	0.3	2.2	0.6		2.2	2.2
Δ moy. dupliqués	0.0	0.0	0.0	0.5	0.1	0.3	0.4	2.0	0.1	1.5	0.1	0.0	0.1	0.1	2.0	0.0		0.4	0.1
TRISTESSE	e1	e2	e3	e4	e5	e6	e7	e8	e9	e10	e11	e12	e13	e14	e15	e16		moy.	méd.
Δ moy. autres	4.2	4.1	3.8	3.9	3.4	3.3	2.8	4.1	3.8	3.5	3.5	4.0	4.0	3.9	3.6	1.2		3.6	3.8
Δ moy. dupliqués	0.4	1.0	3.1	1.0	0.7	0.1	1.1	1.1	0.1	1.5	0.1	1.0	1.0	0.2	2.5	0.4		1.0	1.0

juges v1-v9 | juges v10-v18 | juges e1-e16

*Les distances moyennes entre les expressions dupliquées sont calculées sur la base de 2 valeurs, alors que les distances moyennes entre toutes les autres expressions sont calculées sur la base de 210 valeurs, elles sont donc beaucoup plus stables. En conséquence, les deux types de moyennes ne peuvent être comparées qu'avec réserve.

Relations entre les dimensions vocales et les dimensions émotionnelles évaluées

Des moyennes ont été calculées pour chaque type de jugement (qualités vocales et émotionnelles) et pour chaque expression (32 expressions évaluées). Il importe de relever que dans ce cas les jugements moyens pour la colère, la tristesse et la joie, les jugements moyens pour la hauteur, l'intensité et l'intonation, et les jugements moyens pour l'articulation, la fluidité et la rapidité ont été obtenus sur la base des réponses de 3 groupes d'auditeurs différents. Il existe donc une dépendance entre les dimensions évaluées par un même groupe (par exemple la fluidité, la rapidité et la colère) qui n'existe pas pour les dimensions évaluées par des groupes indépendants (par exemple la tristesse et la rapidité).

Les corrélations entre ces jugements moyens sont représentés dans le tableau 4. On observe que les dimensions vocales sont fortement corrélées entre elles à l'exception de la dimension de 'fluidité' qui ne corrèle avec aucune autre dimension vocale ou émotionnelle. En ce qui concerne les jugements émotionnels, les jugements de tristesse corrèlent avec toutes les dimensions vocales évaluées, alors que les jugements de joie et de peur ne sont corrélés qu'à une partie seulement des dimensions vocales. Les corrélations très importantes qui apparaissent pour les dimensions vocales évaluées par le premier groupe d'auditeur (hauteur, intensité, intonation) pourraient s'expliquer aussi bien par la dépendance des jugements effectués par les mêmes auditeurs que par une relation importante qui existerait entre ces dimensions sur le plan de jugements effectués et/ou dans les expressions évaluées. L'existence de corrélations plus fortes entre certaines dimensions évaluées par des groupes différents, relativement aux corrélations entre des dimensions évaluées par un même groupe, laisse penser que les corrélations entre les différentes dimensions sont d'avantage le produit d'une relation entre ces dimensions que la conséquence d'un jugement répété. Les jugements moyens de tristesse perçue corrèlent par exemple d'avantage avec les jugements de qualité vocale (effectués par des groupes indépendants) qu'avec les jugements émotionnels (effectués par le même groupe d'auditeur).

Tableau 4: Corrélations entre les dimensions (vocales et émotionnelles) évaluées ; les corrélations significatives ($p<.05$) sont représentées en gras.

		hauteur	intensité	intonat.	articulat.	fluidité	rapidité	colère	joie
intensité	r	**0.81**							
intonation	r	**0.79**	**0.65**						
articulation	r	**-0.57**	**-0.55**	**-0.61**					
fluidité	r	-0.23	-0.29	-0.03	-0.02				
rapidité	r	**0.59**	**0.67**	**0.42**	**-0.53**	0.25			
colère	r	0.28	**0.64**	0.09	**-0.38**	-0.31	**0.43**		
joie	r	**0.51**	0.32	**0.64**	-0.27	0.08	0.28	-0.26	
tristesse	r	**-0.51**	**-0.64**	**-0.63**	**0.64**	-0.11	**-0.75**	**-0.41**	**-0.44**

Relations entre les caractéristiques vocales perçues et émotions exprimées

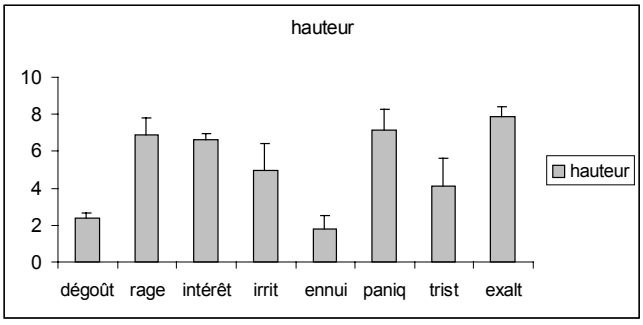
Le nombre d'expressions utilisées (4 par type d'émotion exprimée, 2 pour chaque locuteur) est trop faible pour tester l'effet de l'influence de l'émotion exprimée en tenant compte du locuteur. Les deux "phrases" (énoncés constitués de suite de syllabes sans signification) prononcées ne peuvent pas d'avantage être prises en compte, elles ne sont pas réparties également entre les différentes émotions exprimées. Des ANOVAs effectuées dans ce contexte pour tester l'effet de l'émotion exprimée sur les jugements moyens obtenus pour les expressions sur les différentes dimensions vocales indiquent que l'émotion exprimée affecte significativement les jugements moyens pour toutes les dimensions évaluées, à l'exception de la fluidité. Les résultats de ces ANOVAs sont présentés dans le tableau 5.

Tableau 5: Résultats des ANOVAs, effets de l'émotion exprimée sur les jugements moyens obtenus pour les dimensions vocales évaluées.

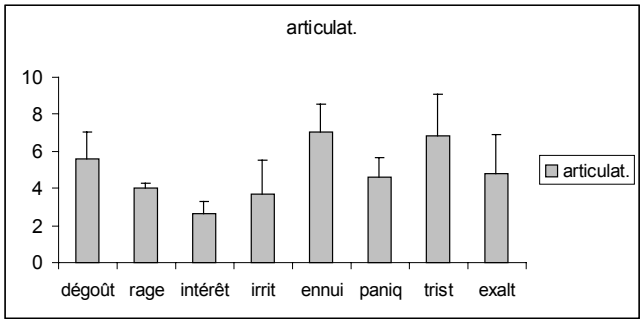
dimension	df	F	Sig.	Eta^2	dimension	df	F	Sig.	Eta^2
hauteur	(7, 24)	22.41	.000	0.87	articulation	(7, 24)	4.07	.004	0.54
intensité	(7, 24)	31.01	.000	0.90	fluidité	(7, 24)	2.06	.088	0.38
intonation	(7, 24)	11.36	.000	0.77	rapidité	(7, 24)	6.89	.000	0.67

Les graphiques 1 à 6 représentent les moyennes et les écarts-types par émotion exprimée (N = 4) pour chacune des 6 dimensions vocales évaluées. Le label 'rage' correspond à la colère chaude, le label 'irrit' correspond à la colère froide, le label 'paniq' correspond à la peur panique, le label 'trist' correspond à la tristesse et le label 'exalt' correspond à la joie intense exprimée.

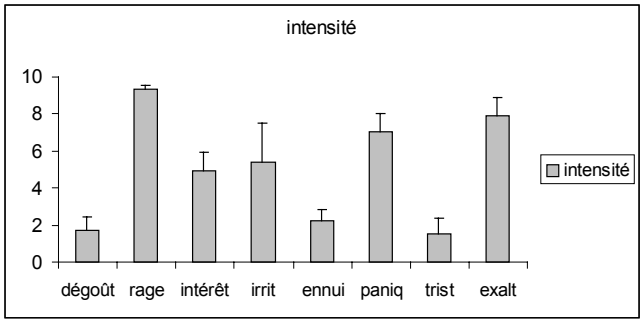
Graphique 1: Degré de hauteur perçue (grave-aigu) en fonction de l'émotion exprimée (N=4)



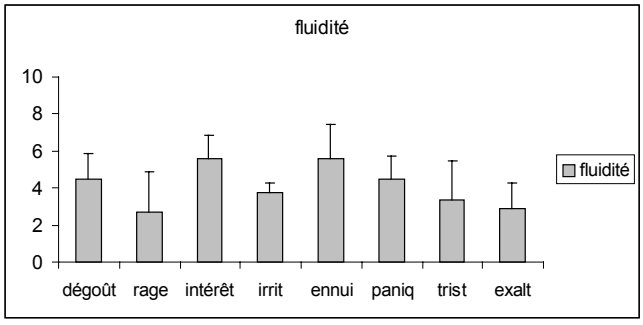
Graphique 2: Qualité de l'articulation perçue (bonne-mauvaise) en fonction de l'émo. expr. (N=4)



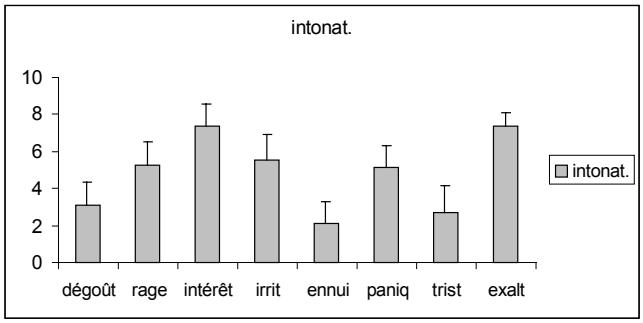
Graphique 3: Degré d'intensité perçue (faible-forte) en fonction de l'émotion exprimée (N=4)



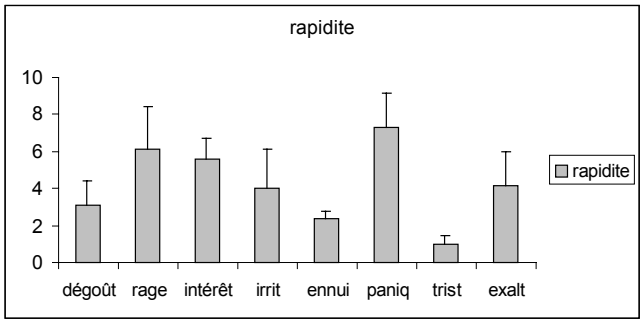
Graphique 4: Degré de fluidité perçue (hâché-fluide) en fonction de l'émotion exprimée (N=4)



Graphique 5: Qualité de l'intonation perçue (monotone-modulée) en fonction de l'émo. exp. (N=4)



Graphique 6: Degré de rapidité perçue (lent-rapide) en fonction de l'émotion exprimée (N=4)



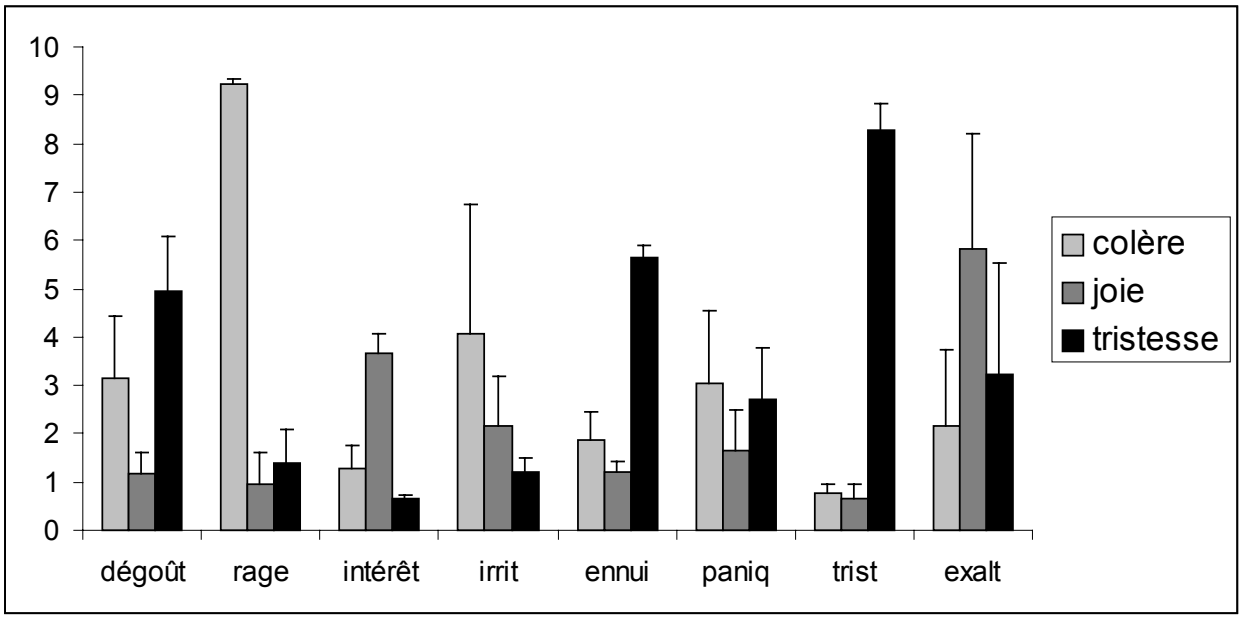
Relations entre les émotions exprimées et les émotions perçues

Ce prétest avait pour but d'évaluer la procédure de jugement pour les caractéristiques vocales qui sont réputées être difficilement évaluables par des auditeurs non-spécialistes. D'autre part nous avons choisi d'utiliser également cette procédure pour obtenir des jugements sur les caractéristiques émotionnelles qui sont habituellement obtenus par d'autres méthodes et qui sont réputées être plus faciles à évaluer. La comparaison entre les jugements obtenus pour les dimensions émotionnelles et pour les dimensions vocales a permis d'établir que les jugements émotionnels ne sont pas plus fidèles que les jugements obtenus pour les caractéristiques vocales en utilisant cette procédure. Dans ce qui suit, une brève description de la relation entre les émotions exprimées dans les enregistrements utilisés et les trois dimensions émotionnelles évaluées est présentée afin d'évaluer la cohérence des jugements émotionnels. Les expressions qui ont été sélectionnées pour le prétest devraient être relativement bien discriminées par les auditeurs et en conséquence devraient recevoir des jugements moyens de tristesse, colère et joie cohérents avec l'émotion exprimée. Le graphique 7 représente les moyennes des jugements pour l'intensité de colère perçue (barres claires), pour l'intensité de joie perçue (barres foncées) et pour l'intensité de tristesse perçue (barres noires) en fonction de l'émotion exprimée. Chacune de ces moyennes est donc calculée à partir des jugements fournis par les groupes d'auditeurs pour 4 expressions émotionnelles (2 expressions * 2 locuteurs).

Les 4 enregistrements qui expriment la colère chaude ('rage') ont reçu un jugement moyen d'intensité émotionnel très élevé pour la colère perçue et très faible pour la joie et la tristesse perçue. Les 4 enregistrements qui expriment la tristesse ('trist') obtiennent un jugement moyen très élevé pour l'intensité de tristesse perçue et des jugements moyens très faibles pour la colère et la joie perçue. En revanche, la différence est beaucoup moins marquée pour les 4 expressions de colère froide ('irrit') qui reçoivent un jugement moyen de colère perçue relativement faible. Cette observation est due essentiellement au fait que les deux expressions de colère froide sélectionnées pour le locuteur 12 n'ont pas été évaluées comme exprimant de la colère (pour K11212, l'intensité

moyenne de colère perçue est égale à 1.7, sd = 1.8 ; pour K21212, l'intensité moyenne de colère perçue est égale à 1.9, sd = 2.4). De même, une expression de joie intense produite par le locuteur 3 n'a pas été identifiée comme exprimant une intensité de joie importante (pour U22203, l'intensité moyenne de joie perçue est égale à 2.4, sd = 3.4).

Graphique 7: Moyennes et écarts-types pour les jugements émotionnels en fonction des émotions exprimées (N = 4)



Conclusions

Dans l'ensemble les résultats du prétest de la procédure de jugement sont satisfaisants, la fidélité inter-juges des réponses est satisfaisante et les jugements présentent une assez bonne cohérence. Les fortes corrélations qui apparaissent entre les dimensions vocales peuvent s'expliquer à la fois par la dépendance qui existe entre ces dimensions dans les expressions présentées (les expressions globalement plus aiguës sont par exemple aussi globalement plus intenses, plus rapides etc…) et/ou par une dépendance qui résulterait de la difficulté qu'auraient les auditeurs à extraire des dimensions vocales séparées à partir des expressions présentées. Bien que fortement corrélées entre elles, les dimensions vocales perçues semblent parvenir à différencier partiellement les émotions exprimées et les émotions perçues.

La dimension de fluidité qui ne corrèle avec aucune autre dimension évaluée et qui présente des indices de fiabilité faibles ne sera pas utilisée pour l'évaluation de la totalité des expressions émotionnelles étudiées. La présence d'une dimension orthogonale aux autres dimensions est en principe souhaitable (la colinéarité entre les dimensions vocales perçues constitue un inconvénient certain), mais les jugements de fluidité sont également indépendants des jugements relatifs aux émotions perçues et indépendants des émotions exprimées. Ils ne présentent donc pas d'intérêt

relativement à l'objectif des études (présentées ci-dessous) qui visent à identifier les relations entre les caractéristiques vocales perçues et les caractéristiques émotionnelles exprimées et perçues.

3.2.3 Procédure pour l'étude principale

La procédure décrite à la section 3.2.1 a été utilisée pour l'étude de jugement principale. Des précisions concernant l'utilisation de cette procédure dans ce contexte spécifique sont présentées dans les sections qui suivent.

3.2.3.1 Enregistrements utilisés

Les 144 enregistrements utilisés dans l'étape 1 (caractéristiques acoustiques) ont été jugés sur plusieurs dimensions vocales. Pour mémoire, ces enregistrements sont produits par 9 acteurs/trices différents qui prononcent deux "phrases" sans signification (1 - "hätt san dik prong nju ven tsie", 2 - "fi gött leich jean kill gos terr") et expriment 8 types d'émotions : colère chaude ('rage') et colère froide ('irrit'), anxiété ('anx') et peur panique ('paniq'), tristesse ('trist') et désespoir ('desp'), joie calme ('joie') et joie intense ('exalt'). Les exemples destinés à uniformiser la compréhension des dimensions vocales et les standards de comparaison interindividuels ont été enregistrés par un locuteur qui ne fait pas partie des locuteurs dont les expressions sont évaluées. Il prononce à chaque fois la phrase « je ne peux pas le croire ».

3.2.3.2 Auditeurs, dimensions vocales et design de l'étude de jugement

La procédure de jugements décrite à la section 3.2.1 étant plus coûteuse en temps qu'une procédure conventionnelle, quatre groupes d'auditeurs ont été recrutés. Les groupes sont composés chacun de 15 à 16 étudiants de 1ère année en Psychologie à l'Université de Genève. Ces étudiants ont participé à cette étude en échange d'un crédit de cours. Ils ont été affectés aléatoirement aux 4 groupes, ils sont tous francophones et ne souffrent pas de déficits auditifs diagnostiqués. Chaque groupe a évalué les expressions produites par 3 locuteurs différents. Afin de tester l'absence de différences systématiques entre les groupes, les enregistrements produits par un locuteur sélectionné aléatoirement (l'acteur 3) ont été évalués par les 4 groupes. Un premier groupe constitué de 14 femmes et 2 hommes (âge moyen = 21.3 ans, sd = 4.3) a évalué les expressions produites par l'actrice 2, l'acteur 3 et l'acteur 4 ; un deuxième groupe composé de 10 femmes et 5 hommes (âge moyen = 21.7, sd = 4.3) a jugé les expressions produites par l'actrice 7, l'acteur 8 et l'acteur 3 ; un troisième groupe composé de 13 femmes et 2 hommes (âge moyen = 20.2, sd = 1.7) a évalué les expressions produites par l'actrice 9, l'actrice 10 et l'acteur 3 ; et un quatrième groupe composé de 11 femmes et 4 hommes (âge moyen = 21.4, sd = 6.1) a évalué les expressions produites par l'actrice 11, l'acteur 12 et l'acteur 3.

Chaque auditeur a évalué, en deux sessions de 1h30 à une semaine d'intervalle, 48 (2 phrases x 8 émotions exprimées x 3 locuteurs) enregistrements sur les 8 dimensions suivantes: 'hauteur' (*hauteur*, voix *grave* ↔ *aiguë*), 'intensité' (*volume*, voix *faible* ↔ *forte*), 'intonation' (*mélodie*, voix *monotone* ↔ *modulée*), 'rapidité' (*vitesse*, *lente* ↔ *rapide*), 'articulation' (*articulation*, *mal articulée* ↔ *bien articulée*), 'instabilité' (*stabilité*, *ferme* ↔ *tremblante*), qualité 'rauque' (*voix rauque*, *pas rauque* ↔ *rauque*) et 'perçante' (*voix perçante*, *pas perçante* ↔ *perçante*). Les termes entre parenthèses et en italique correspondent aux termes présentés aux participants sur l'écran de l'ordinateur durant les évaluations.

Les expressions et les dimensions ont été présentées selon le schéma suivant: Une moitié des auditeurs de chaque groupe, sélectionnée aléatoirement, a évalué les dimensions 'hauteur', 'intensité', 'instabilité', 'rauque' durant la première session et les dimensions 'intonation', 'rapidité', 'articulation', 'perçante' durant la deuxième session. L'autre moitié des auditeurs a évalué d'abord les dimensions 'intonation', 'rapidité', 'articulation', 'perçante' et ensuite les dimensions 'hauteur', 'intensité', 'instabilité', 'rauque'. Pour chaque auditeur, un ordre aléatoire de présentation des expressions est défini par l'ordinateur au début de chaque session. Un locuteur est sélectionné aléatoirement parmi les 3 locuteurs que chaque auditeur est chargé d'évaluer. Les expressions du premier locuteur sélectionné sont évaluées successivement sur les quatre dimensions prédéfinie pour chaque session. L'ordre de présentation de ces dimensions est défini aléatoirement par l'ordinateur. Un deuxième locuteur est ensuite choisi aléatoirement parmi les 2 locuteurs restant et un nouvel ordre de présentation aléatoire des 4 dimensions est défini. Finalement, un troisième ordre aléatoire de présentation des dimensions est défini pour l'évaluation des expressions produites par le dernier locuteur. D'autre part, pour chaque dimension présentée, l'ordinateur définit un nouvel ordre aléatoire de présentation des expressions produites par le locuteur évalué. Pour rappel, les expressions produites par un même locuteur sont présentées successivement, mais sont finalement disponibles simultanément sur chaque dimension.

3.3 Résultats

3.3.1 Evaluation des différences de réponse entre les groupes d'auditeurs

Les expressions produites par le locuteur 3 (16 expressions) ont été évaluées par l'ensemble des auditeurs afin de permettre de vérifier pour ces expressions que les jugements produits par les auditeurs ne sont pas dépendants des 4 groupes d'auditeurs créés. Huit ANOVAs à mesures répétées (avec un facteur 'expression' intra/within à 16 niveaux et un facteur 'groupe' inter/between à 4 niveaux) ont été effectuées pour tester l'existence d'un éventuel effet (inter/between) du groupe d'auditeurs sur les jugements effectués pour chacune des 8 dimensions vocales. L'effet principal du

groupe est non-significatif pour 7 dimensions vocales. Sa taille varie de $\eta^2 = 0.01$ pour les jugements relatifs à l'intensité perçue à $\eta^2 = 0.11$ pour les jugements relatifs à la qualité de l'articulation. L'effet principal du groupe est significatif pour les jugements sur la dimension vocale 'perçante': $F(3, 57) = 3.49$, $p = .021$, $\eta^2 = 0.16$. La taille de cet effet est toutefois relativement faible en comparaison avec l'effet (inta/within) des expressions sur ces jugements: $F(15, 855) = 3.49$, $p < .001$, $\eta^2 = 0.56$. Selon le test post-hoc HSD de Tukey, l'effet du groupe sur les jugements relatifs à la qualité perçante est attribuable au fait que les jugements du 4ème groupe de juges – qui évalue par ailleurs les expressions produites par les locuteurs 11 et 12 – sont significativement plus élevés (moy = 5.8) pour cette dimension que les jugements du 2ème groupe de juges (moy = 4.9) qui évalue par ailleurs les expressions produites par les locuteurs 7 et 8. L'interaction entre le facteur 'expression' (16 niveaux, intra) et le facteur 'groupe' (4 niveaux, inter) est non-significative pour les 8 dimensions vocales évaluées, alors que les expressions ont un effet principal significatif sur les jugements pour l'ensemble de ces dimensions.

Un effet systématique du groupe n'est donc pas présent. L'unique différence significative entre les jugements du 2ème et du 4ème groupe pour la dimension 'perçante' ($p = .032$ selon le test post-hoc) ne permet pas de pronostiquer l'existence d'autres différences entre les groupes de juges formés.

3.3.2 Fidélité inter-auditeurs des jugements

Afin d'évaluer si les moyennes des jugements fournis par les différents groupes d'auditeurs pour chacune des expressions émotionnelles sur les 8 dimensions vocales peuvent être utilisées, des indices de fidélité inter-auditeurs ont été calculés pour chaque dimension vocale et pour chaque groupe d'auditeurs. Le tableau 6 représente les corrélations intraclasses calculées pour les 8 dimensions et pour chaque groupe d'auditeurs. La valeur r correspond à une estimation de la corrélation moyenne entre les jugements de toutes les paires d'auditeurs. La valeur R représente l'indice de fidélité des jugements (équivalent à la valeur α de Cronbach). Pour mémoire, chaque groupe évalue 48 expressions (dont 32 sont différentes), l'effectif de chaque groupe est indiqué entre parenthèses dans le tableau 6.

Dans l'ensemble, les jugements des auditeurs présentent une bonne fidélité. Une seule dimension (qualité 'rauque') évaluée par le groupe 2 présente un coefficient de fidélité inférieur à 0.80. Les jugements d'intensité sont particulièrement fidèles pour les 4 groupes d'auditeurs.

Tableau 6: Coefficients de fidélité (r = single measure intraclass correlation, R = average measure intraclass correlation) pour les 8 dimensions vocales et les 4 groupes d'auditeurs.

dimension	groupe 1 (N=16)		groupe 2 (N=15)		groupe 3 (N=15)		groupe 4 (N=15)	
	r	R	r	R	r	R	r	R
articulation	0.27	0.85	0.31	0.87	0.22	0.81	0.47	0.93
intonation	0.41	0.92	0.34	0.89	0.37	0.90	0.47	0.93
intensité	0.85	0.99	0.84	0.99	0.84	0.99	0.88	0.99
hauteur	0.58	0.96	0.49	0.93	0.51	0.94	0.50	0.94
rauque	0.29	0.87	0.19	0.78	0.30	0.87	0.44	0.92
rapidité	0.64	0.97	0.62	0.96	0.66	0.97	0.72	0.98
perçante	0.39	0.91	0.61	0.96	0.63	0.96	0.72	0.97
instabilité	0.49	0.94	0.50	0.94	0.41	0.91	0.47	0.93

3.3.3 Relations entre les caractéristiques vocales perçues

Une fidélité suffisante des jugements étant établie, des jugements moyens peuvent être calculés pour chacune des 144 expressions émotionnelles et pour chacune des 8 dimensions vocales. Le fait que les expressions produites par l'acteur 3 aient été évaluées par 4 fois plus d'auditeurs que les autres expressions pose dès lors problème. Les moyennes calculées pour les expressions produites par ce locuteur seraient basées sur 4 fois plus de jugements que les moyennes produites par les 8 autres locuteurs. En conséquence, une partie des jugements effectués pour le locuteur 3 ont été supprimés aléatoirement: Les jugements de 4 auditeurs du 1er groupe, de 4 auditeurs du 2ème groupe, de 4 auditeurs du 3ème groupe et de 4 auditeurs du 4ème groupe d'auditeurs ont été sélectionnés aléatoirement pour les expressions produites par le locuteur 3, les jugements des autres auditeurs ont été éliminé du calcul de la moyenne. Les moyennes sont dès lors toutes basées sur 15 ou 16 jugements pour chaque expression (16 jugements pour les expressions des locuteurs 2, 3 et 4; 15 jugements pour les expressions des locuteurs 7, 8, 9, 10, 11 et 12).

Les jugements moyens ainsi définis pour les 144 expressions émotionnelles ont été corrélés pour les 8 dimensions vocales évaluées. Ces corrélations sont représentées dans le tableau 7. On observe que les dimensions sont dans l'ensemble fortement corrélées. Pour les 144 expressions évaluées, les jugements moyens relatifs à la qualité de l'articulation ne corrèlent pas avec les jugements moyens de hauteur et les jugements moyens de rapidité. De même, les jugements relatifs à la qualité rauque sont indépendants des jugements de hauteur et des jugements relatifs à l'intonation perçue ; les jugements d'instabilité ne sont pas liés aux jugements d'intensité, d'intonation et de qualité perçante.

Tableau 7: Corrélations entre les 8 dimensions émotionnelles évaluées, les corrélations significatives ($p < .05$) sont représentées en gras

		hauteur	intensité	intonation	articulation	perçante	rapidité	rauque
intensité	r	**0.67**						
intonation	r	**0.83**	**0.72**					
articulation	r	0.12	**0.30**	**0.34**				
perçante	r	**0.81**	**0.93**	**0.81**	**0.22**			
rapidité	r	**0.51**	**0.66**	**0.55**	0.09	**0.67**		
rauque	r	0.08	**0.42**	0.09	**-0.32**	**0.41**	**0.18**	
instabilité	r	**0.28**	-0.06	0.08	**-0.54**	0.07	**-0.20**	**0.29**

Une analyse en composantes principales indique qu'un peu moins des 3/4 de la variance de ces 8 dimensions peut être expliquée par 2 composantes. Les résultats de cette analyse sont présentés dans le tableau 8 qui représente, à gauche, les valeurs propres et les pourcentages de variance expliquée par les composantes et, à droite, les saturations des dimensions vocales sur les 2 premières composantes de l'ACP. La première composante regroupe des dimensions ('perçante', 'intensité', 'intonation', 'hauteur' et 'rapidité') qui sont en théorie corrélées dans les expressions présentées et qui varieraient en particulier avec le degré d'activation émotionnelle. La deuxième composante regroupe les dimensions 'instabilité', 'articulation' et 'rauque' qui pourraient être liées plutôt avec la "bonne" respectivement la "mauvaise" qualité perçue de la parole et de la voix.

Tableau 8: Résultats de l'analyse en composantes principales: valeurs propres et pourcentages de variance expliquée (partie gauche du tableau), saturations des dimensions vocales sur les 2 premières composantes de l'ACP (partie droite du tableau)

composantes	Valeurs propres	% de variance expliquée	% cumulé
1.00	4.05	50.56	50.56
2.00	1.81	22.64	73.21
3.00	0.99	12.41	85.61
4.00	0.64	7.98	93.59
5.00	0.20	2.55	96.14
6.00	0.17	2.10	98.24
7.00	0.10	1.29	99.53
8.00	0.04	0.47	100.00

dimensions vocales	Compo-sante 1	Compo-sante 2
perçante	0.974	
intensité	0.926	
intonation	0.886	
hauteur	0.851	
rapidité	0.740	
instabilité		0.838
articulation		-0.824
rauque	0.325	0.625

3.3.4 Relations entre caractéristiques vocales perçues et caractéristiques acoustiques

Dans la première étape (caractéristiques acoustiques des expressions), 10 paramètres acoustiques ont été sélectionnés sur la base des résultats d'une analyse en composantes principales et en fonction de leur intérêt théorique (v. section 2.3.2.3). Il s'agissait de: l'intensité moyenne (int.moy), l'étendue de l'intensité (int.étdu), la F0 minimale (F0.min), l'étendue de F0 (F0.étdu), la durée totale (dur.tot),

la durée relative des parties voisées sur les parties voisées et non-voisées (dur.v/art), la proportion d'énergie voisée en dessous de 1000 Hz (v.0-1k), la proportion d'énergie non-voisée en dessous de 1000 Hz (n.0-1k), la proportion d'énergie voisée entre 300 et 500 Hz (v.300-500), la proportion d'énergie voisée entre 600 et 800 Hz (v.600-800). Ces 10 paramètres acoustiques ont été entrés comme prédicteurs des dimensions vocales perçues dans des régressions multiples. Une procédure 'stepwise' a été utilisée afin de définir empiriquement la meilleure solution pour chaque dimension vocale. Le tableau 9 présente les résultats de ces régressions. Pour chaque dimension vocale perçue, le modèle qui permet d'expliquer le plus de variance est présenté. On constate que différentes combinaisons de paramètres acoustiques parviennent à rendre compte d'une très large partie de la variance de l'intensité vocale perçue (88%) et de la qualité perçante perçue (87%). Les paramètres acoustiques utilisés expliquent également une assez large part de la variance des jugements moyens de rapidité (79%), d'intonation (67%) et de hauteur (65%) et une part plus faible de la variance des jugements moyens d'instabilité (35%), de la qualité de l'articulation (32%) et de la qualité rauque (28%).

Tableau 9: Régressions (stepwise) des 10 paramètres acoustiques sur les dimensions vocales.

dimension vocale	param. acoust.	b	t	sig.	dimension vocale	param. acoust.	b	t	sig.
INTENSITÉ	int.moy	0.72	15.97	.000	HAUTEUR	F0.min	0.41	7.11	.000
	int.étdu	0.18	5.50	.000		F0.étdu	0.40	6.25	.000
$R^2 = 0.88$	v.0-1k	-0.16	-3.76	.000	$R^2 = 0.65$	int.moy	0.25	3.49	.001
RAPIDITÉ	dur.tot	-0.57	-13.32	.000	RAUQUE	n.0-1k	0.26	3.49	.001
	int.moy	0.43	6.28	.000		v.0-1k	-0.35	-3.27	.001
	v.0-1k	-0.18	-3.11	.002		int.moy	0.23	2.09	.038
$R^2 = 0.79$	dur.v/art	-0.15	-2.97	.003	$R^2 = 0.28$	dur.tot	0.22	2.83	.005
PERÇANTE	int.moy	0.51	9.14	.000	ARTICULATION	F0.min	-0.39	-4.78	.000
	F0.étdu	0.25	6.28	.000		int.moy	0.47	4.56	.000
	F0.min	0.18	5.07	.000		int.étdu	0.31	3.76	.000
	v.0-1k	-0.15	-3.28	.001		F0.étdu	-0.30	-3.29	.001
$R^2 = 0.87$	int.étdu	0.08	2.11	.036	$R^2 = 0.32$	n.0-1k	-0.20	-2.71	.008
INSTABILITÉ	F0.min	0.49	6.56	.000	INTONATION	F0.étdu	0.40	5.85	.000
	dur.tot	0.35	4.88	.000		int.moy	0.21	2.51	.013
$R^2 = 0.35$	v.0-1k	0.24	3.07	.003		int.étdu	0.21	3.53	.001
						F0.min	0.19	3.25	.001
						n.0-1k	-0.15	-2.84	.005
					$R^2 = 0.67$	dur.tot	-0.13	-2.23	.027

Les paramètres acoustiques sélectionnés parviendraient donc à bien rendre compte des jugements pour les dimensions qui saturent sur la première composante de l'ACP (tableau 8) et dans une moindre mesure des jugements qui saturent sur la deuxième composante de l'ACP. La régression (enter) des paramètres acoustiques sur les scores factoriels de la première et de la deuxième

composante de l'ACP confirment cette impression. Les 10 paramètres acoustiques expliquent conjointement 88% de la variance de la première composante de l'ACP et ne parviennent à rendre compte que de 34% de la variabilité de la deuxième composante de l'ACP.

Aucun des paramètres acoustiques mesurés ne correspondent directement aux dimensions vocales évaluées. Les corrélats acoustiques exacts de ces dimensions sont très difficiles à définir pour des enregistrements de parole émotionnelle. Toutefois, certains paramètres acoustiques sont théoriquement plus directement liés aux dimensions vocales perçues que d'autres paramètres. L'intensité moyenne (int.moy) par exemple pourrait être un relativement bon prédicteur de l'intensité perçue. Le tableau 10 présentent quelques corrélations entre les dimensions vocales perçues 'intensité', 'hauteur', 'rapidité', 'intonation', 'perçante' et 5 paramètres vocaux sélectionnés pour leur proximité conceptuelle avec chacune de ces dimensions. Des paramètres acoustiques qui seraient théoriquement des prédicteurs aussi directs des dimensions 'articulation', 'instabilité' et 'rauque' n'ont pas été mesurés.

Tableau 10: Corrélations entre 5 dimensions vocales perçues et 5 paramètres acoustiques conceptuellement liés

	intonation	intensité	hauteur	rapidité	perçante
F0.étdu	**0.63**	0.60	0.61	0.33	0.67
int.moy	0.71	**0.92**	0.68	0.70	0.89
F0.min	0.43	0.46	**0.57**	0.33	0.53
dur.tot	-0.24	-0.27	-0.17	**-0.74**	-0.27
v.0-1k	-0.56	-0.74	-0.46	-0.60	**-0.72**

Dans une conception naïve de la correspondance entre les paramètres acoustiques et les dimensions vocales perçues, les corrélations représentées en gras (diagonale du tableau 10) devraient être plus élevées que les autres corrélations figurant dans la même colonne. On observe que cela n'est pas vrai pour toutes les dimensions vocales perçues. L'intensité acoustique moyenne semble être un meilleur prédicteur de l'intonation perçue, de la hauteur perçue et de la qualité perçante perçue que, respectivement, l'étendue de F0, la F0 minimale et la proportion d'énergie voisée dans les basses fréquences. La comparaison des corrélations présentées dans le tableau 10 et des régressions multiples présentées dans le tableau 9 met en évidence un aspect problématique de la colinéarité des données présentées. L'intensité acoustique moyenne, en particulier, corrèle assez fortement avec l'ensemble des autres paramètres acoustiques utilisés. Lorsqu'elle est associée à ces paramètres dans les régressions multiples, des coefficients plus importants apparaissent pour des paramètres acoustiques qui, lorsqu'ils sont considérés de manière isolée, ont une corrélation plus faible que l'intensité acoustique moyenne avec les dimensions vocales perçues. Par exemple, pour la prédiction de la variance de la hauteur perçue, les contributions indépendantes respectives de la F0 minimale et de l'étendue de F0 (utilisées en association avec l'intensité acoustique moyenne comme

prédicteurs de la hauteur perçue) sont supérieures à la contribution indépendante de l'intensité acoustique moyenne.

3.3.5 Relations entre caractéristiques vocales perçues et émotions exprimées

Les relations entre les 8 types d'émotions exprimées et les jugements moyens obtenus pour les expressions sur les 8 dimensions vocales perçues ont été testées en effectuant 8 ANOVAs à mesures répétées avec les facteurs intra (within) 'phrase' (2 niveaux) et 'émotion exprimée' (8 niveaux). Pour les 8 dimensions vocales, l'effet de l'émotion exprimée est significatif alors que l'effet de la phrase et de l'interaction entre émotion et phrase est non-significatif. La taille des effets de l'émotion exprimée sur les dimensions vocales perçues varie de $\eta^2 = 0.44$ pour la qualité rauque perçue à $\eta^2 = 0.82$ pour la qualité perçante perçue. Les résultats de ces ANOVAs sont présentés dans le tableau 11.

Tableau 11: ANOVAs à mesures répétées - Effets "within" de la phrase et de l'émotion sur les jugements moyens obtenus pour les 8 dimensions vocales perçues, les résultats significatifs ($p<.05$) sont indiqués en gras.

dimension	source	df	F	sig.	eta^2	dimension	source	df	F	sig.	eta^2
intensité	phrase	(1, 8)	0.00	.969	0.00	perçante	phrase	(1, 8)	0.12	.738	0.01
	emo	**(7, 56)**	**32.54**	**.000**	**0.80**		**emo**	**(7, 56)**	**36.00**	**.000**	**0.82**
	phrase*emo	(7, 56)	1.64	.143	0.17		phrase*emo	(7, 56)	1.19	.323	0.13
hauteur	phrase	(1, 8)	0.07	.792	0.01	articulation	phrase	(1, 8)	0.71	.423	0.08
	emo	**(7, 56)**	**19.85**	**.000**	**0.71**		**emo**	**(7, 56)**	**7.87**	**.000**	**0.50**
	phrase*emo	(7, 56)	1.30	.267	0.14		phrase*emo	(7, 56)	1.42	.218	0.15
intonation	phrase	(1, 8)	0.05	.836	0.01	rauque	phrase	(1, 8)	4.31	.071	0.35
	emo	**(7, 56)**	**32.15**	**.000**	**0.80**		**emo**	**(7, 56)**	**6.26**	**.000**	**0.44**
	phrase*emo	(7, 56)	0.36	.921	0.04		phrase*emo	(7, 56)	0.85	.551	0.10
rapidité	phrase	(1, 8)	0.08	.790	0.01	instabilité	phrase	(1, 8)	0.59	.464	0.07
	emo	**(7, 56)**	**20.02**	**.000**	**0.71**		**emo**	**(7, 56)**	**15.87**	**.000**	**0.66**
	phrase*emo	(7, 56)	0.44	.871	0.05		phrase*emo	(7, 56)	0.89	.522	0.10

Ces ANOVAs ont par ailleurs fait apparaître un effet inter (between) significatif pour les 8 dimensions vocales perçues. La présence de différences entre les locuteurs n'est en principe pas surprenante pour ces dimensions: on peut s'attendre, par exemple, à ce que les voix féminines soient perçues comme plus aiguës que les voix masculines. Cependant, la procédure utilisée, dans laquelle les expressions produites par chaque locuteur sont évaluées séparément, aurait en théorie dû contrôler une large partie de ces différences. La présence d'un effet assez important du locuteur pourrait donc signifier, contre nos attentes, que les auditeurs auraient utilisé un même standard de comparaison pour les présentations répétées d'une même dimension. L'effet inter (between) significatif dans les ANOVAs à mesures répétées pourrait également être dû à des effets d'interaction entre les locuteurs et les émotions exprimées. Afin d'évaluer la nature de la contribution du locuteur à la variance des dimensions vocales perçues, 8 ANOVAs ont été

effectuées avec 2 facteurs fixes: 'phrase' (2 niveaux) et 'émotion exprimée' (8 niveaux) et un facteur random: 'locuteur' (9 niveaux). Dans ces analyses, les effets principaux de l'émotion exprimée, de la phrase et les effets d'interaction entre l'émotion exprimée et la phrase sont identiques aux effets rapportés dans le tableau 11 pour les ANOVAs à mesures répétées. Les 8 ANOVAs effectuées indiquent également les effets principaux du locuteur (facteur random) et les effets d'interaction entre le locuteur et la phrase, ainsi que les effets d'interaction entre le locuteur et l'émotion exprimée. Ces effets sont rapportés dans le tableau 12. Les effets principaux du locuteur ne sont pas significatifs et leur taille est relativement faible, alors que les effets d'interaction entre l'émotion exprimée et le locuteur sont significatifs et plus importants.

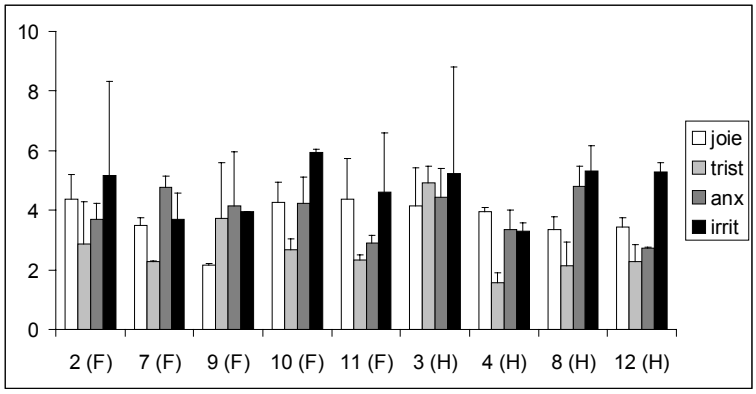
Tableau 12: ANOVAs - Effets "random" du locuteur et effets d'interactions avec les émotions exprimées et les phrases sur les jugements moyens obtenus pour les 8 dimensions vocales perçues, les résultats significatifs ($p<.05$) sont indiqués en gras.

dimension	source	df	F	sig.	eta^2	dimension	source	df	F	sig.	eta^2
intensité	locuteur	(8, 26)	1.41	.238	0.31	perçante	locuteur	(8, 15)	0.77	.632	0.29
	emo * locut.	**(56, 56)**	**2.69**	**.000**	**0.73**		**emo * locut.**	**(56, 56)**	**2.31**	**.001**	**0.70**
	phrase * locut.	(8, 56)	1.18	.326	0.14		**phrase * locut.**	**(8, 56)**	**2.95**	**.008**	**0.30**
hauteur	locuteur	(8, 14)	0.54	.809	0.23	articulation	locuteur	(8, 22)	0.84	.578	0.23
	emo * locut.	**(56, 56)**	**2.07**	**.004**	**0.67**		**emo * locut.**	**(56, 56)**	**2.32**	**.001**	**0.70**
	phrase * locut.	**(8, 56)**	**2.61**	**.017**	**0.27**		phrase * locut.	(8, 56)	0.99	.450	0.12
intonation	locuteur	(8, 12)	0.32	.943	0.18	rauque	locuteur	(8, 31)	1.09	.397	0.22
	emo * locut.	**(56, 56)**	**1.86**	**.011**	**0.65**		**emo * locut.**	**(56, 56)**	**3.48**	**.000**	**0.78**
	phrase * locut.	**(8, 56)**	**3.59**	**.002**	**0.34**		phrase * locut.	(8, 56)	1.46	.191	0.17
rapidité	locuteur	(8, 12)	0.53	.812	0.26	instabilité	locuteur	(8, 22)	0.53	.818	0.17
	emo * locut.	**(56, 56)**	**1.82**	**.014**	**0.65**		**emo * locut.**	**(56, 56)**	**4.20**	**.000**	**0.81**
	phrase * locut.	**(8, 56)**	**2.78**	**.012**	**0.28**		**phrase * locut.**	**(8, 56)**	**4.19**	**.001**	**0.37**

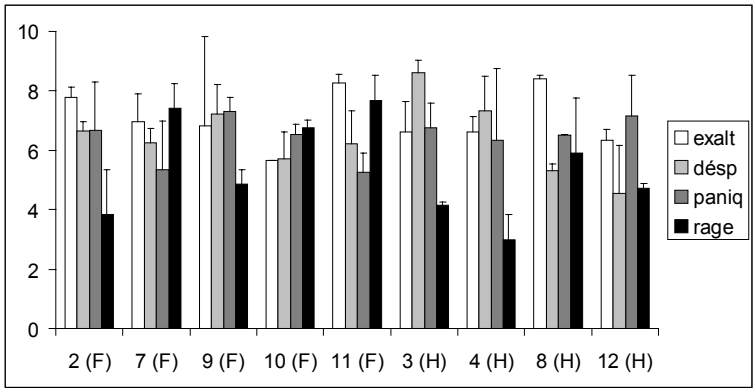
Les graphiques 8 à 10 illustrent cette observation pour la hauteur perçue. Des effets d'interaction très marqués apparaissent sur les graphiques 8 et 9, alors que la hauteur perçue ne varie pas en fonction directe du locuteur (graphique 10). Les moyennes représentées par les barres dans les graphiques 8 et 9 sont basées sur 2 observations seulement, elles ne peuvent donc être généralisées.

Les résultats présentés dans le tableau 12 et les graphiques 8 à 10 démontrent que les effets attribuables aux locuteurs ne sont pas dus à un standard de comparaison stable relativement auquel les expressions des différents locuteurs seraient évaluées (dans ce cas, les expressions produites par les femmes auraient probablement été jugées plus aiguës que les expressions produites par les hommes). Au contraire, les expressions produites par un même locuteur semblent bien avoir été jugées relativement à la variabilité de ce locuteur.

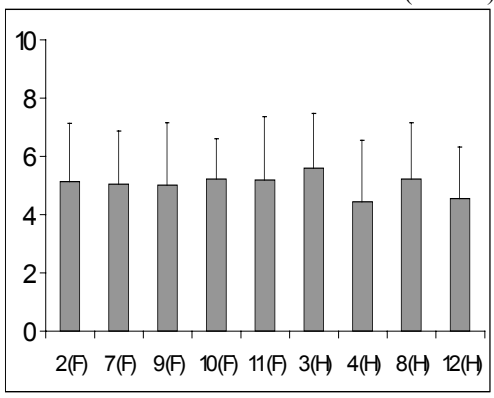
Graphique 8: Hauteur moyenne perçue en fonction du locuteur et de l'émotion exprimée pour les émotions faiblement activées (N = 2)



Graphique 9: Hauteur moyenne perçue en fonction du locuteur et de l'émotion exprimée pour les émotions fortement activées (N = 2)



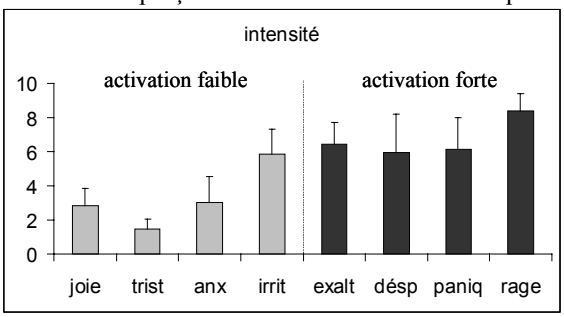
Graphique 10: Hauteur moyenne perçue en fonction du locuteur (N = 16)



Les lettres entre parenthèses indiquent le sexe du locuteur: F = femme, H = homme.

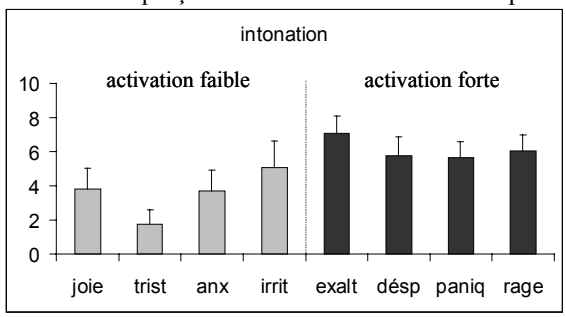
Les graphiques 11 à 18 représentent les moyennes et les écarts-types des jugements pour chaque dimension évaluée en fonction des émotions exprimées. Les moyennes et les écarts-types présentés dans ces graphiques sont calculés sur la base de 18 expressions (9 locuteurs * 2 phrases) et sont également représentés en annexe B2. On observe des différences importantes pour les émotions exprimées sur les 8 dimensions évaluées. Les différences significatives selon le test post-hoc HSD de Tukey pour une série de 8 ANOVAs sont rapportées sous les graphiques. Les graphiques 11 à 18 indiquent également qu'un ensemble différent de propriétés vocales perçues pourrait caractériser chaque émotion exprimée. Les expressions de tristesse, par exemple, sont perçues comme moins intenses, plus monotones et plus lentes, alors que les expressions de colère froide (irrit) se caractérisent par une voix jugée moins tremblante et une articulation jugée meilleure que la plupart des autres expressions.

Graphique 11: Moyennes et écarts-types pour l'intensité perçue en fonction de l'émotion exprimée



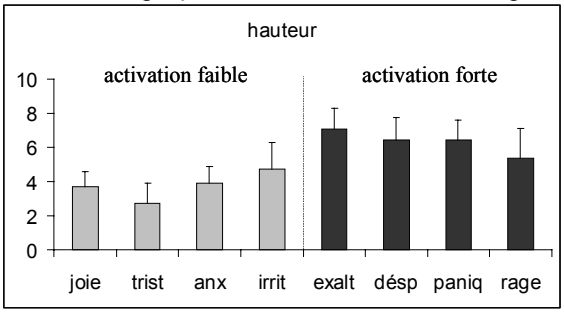
- trist < anx, irrit, désp, paniq, exalt, rage
- joie, anx < irrit, désp, paniq, exalt, rage
- irrit, désp, paniq, exalt < rage

Graphique 12: Moy. et écarts-types pour l'intonation perçue en fonction de l'émotion exprimée



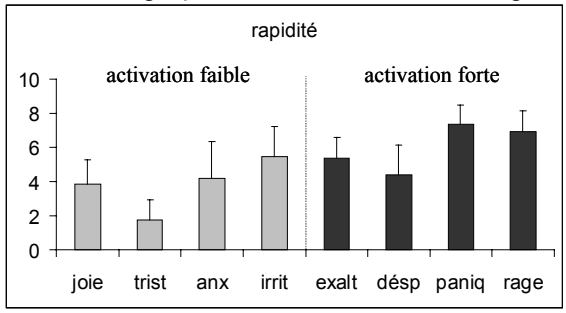
- trist < anx, joie, irrit, paniq, désp, rage, exalt
- anx, joie < irrit, paniq, désp, rage, exalt
- irrit, paniq, désp < exalt

Graphique 13: Moyennes et écarts-types pour la hauteur perçue en fonction de l'émotion exprimée



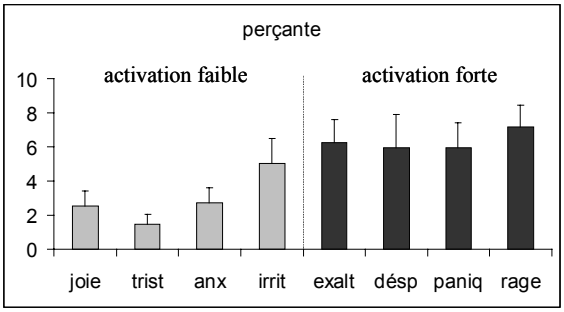
- trist < irrit, rage, desp, paniq, exalt
- joie, anx < rage, desp, paniq, exalt
- irrit < desp, paniq, exalt
- rage < exalt

Graphique 14: Moyennes et écarts-types pour la rapidité perçue en fonction de l'émotion exprimée



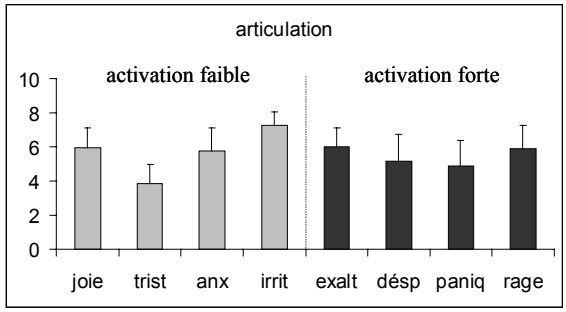
- trist < joie, anx désp, exalt, irrit, rage, paniq
- joie < irrit, rage, paniq
- anx, désp, exalt < rage, paniq
- irrit < paniq

Graphique 15: Moy. et écarts-types pour la qualité perçante en fonction de l'émotion expr.



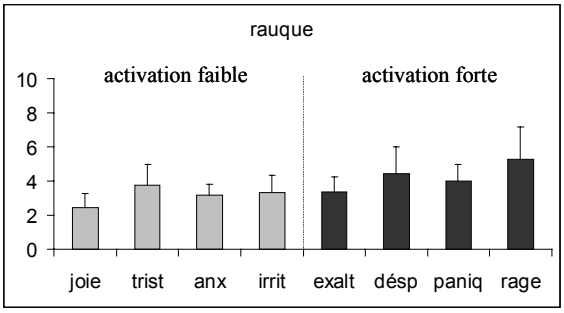
- trist, joie, anx < irrit, désp, paniq, exalt, rage
- irrit < rage

Graphique 16: Moy. et écarts-types pour la qualité de l'articulation en fonction de l'émo. expr.



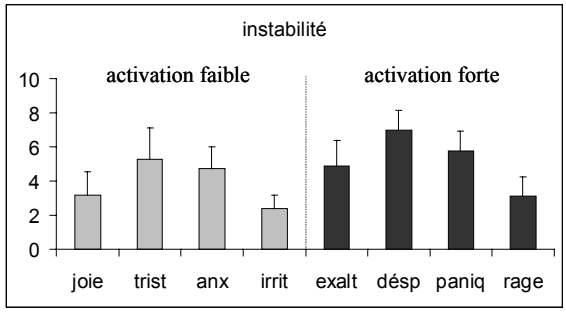
- trist < désp, anx, rage, joie, exalt, irrit
- paniq, désp, anx, rage, joie, exalt < irrit

Graphique 17: Moy. et écarts-types pour la qualité rauque en fonction de l'émotion exprimée



- joie < trist, paniq, désp, rage
- anx < désp, rage irrit, ▪ exalt, trist, paniq < rage

Graphique 18: Moyennes et écarts-types pour l'instabilité perçue en fonction de l'émotion exprimée



- irrit, rage, joie < anx, exalt, trist, paniq, désp
- anx, exalt, trist < désp

La possibilité de discriminer les 8 types d'émotions exprimés par les locuteurs dans cette étude à l'aide des propriétés vocales perçues a été examinée plus globalement en effectuant une analyse discriminante. Les valeurs propres et le pourcentage de variance expliquée par les fonctions extraites par cette analyse sont présentées dans le tableau 12. La dernière colonne de ce tableau fourni une estimation de l'importance de la relation entre chaque fonction et les catégories (émotions exprimées) à discriminer. Selon le lambda de Wilks, les 5 premières fonctions discriminantes apportent une contribution significative à la discrimination.

Tableau 12 : Résultats de l'analyse discriminante: valeurs propres et % de variance expliquée par les fonctions discriminantes, relations entre les groupes et les fonctions (corr. canonique).

Fonction	valeur propre	% de variance	% cumulé de variance	corrélation canonique
1	3.674	57.3	57.3	.887
2	1.536	24.0	81.2	.778
3	.594	9.3	90.5	.610
4	.418	6.5	97.0	.543
5	.150	2.3	99.4	.361
6	.036	.6	99.9	.187
7	.004	.1	100.0	.065

Les 7 fonctions discriminantes dérivées des jugements moyens obtenus pour les 8 dimensions permettent de reclasser 69.4% des expressions dans leurs catégories d'origine. Lorsque les fonctions discriminantes sont calculées pour classer chaque expression en se basant sur l'ensemble des expressions excepté l'expression à classer ("cross validation"), 61.1% des expressions sont encore classées correctement. Le nombre d'expressions classées dans chaque catégorie par cette seconde analyse est représenté dans le tableau 13.

Tableau 13 : Classification des expressions par l'analyse discriminante (cross validated) basée sur les 8 dimensions vocales évaluées.

Groupe original	**Appartenance de groupe prédite**							
	1 (anx.)	2 (joie)	3 (rage)	4 (irrit.)	5 (paniq.)	6 (trist.)	7 (exalt.)	8 (désp.)
Anx.	**6**	3	0	1	2	4	0	2
Joie	4	**10**	0	2	0	2	0	0
Rage	0	0	**12**	3	0	0	2	1
Irrit.	0	2	1	**12**	0	0	3	0
Paniq.	1	0	1	1	**12**	0	1	2
Trist.	0	1	0	0	0	**16**	0	1
Exalt.	0	1	0	2	3	0	**10**	2
Désp.	3	0	0	0	3	1	1	**10**

On notera que la discrimination des émotions exprimées réalisée à l'aide des dimensions vocales perçues a permis de reclasser dans leurs catégories d'origine une proportion plus importante

d'expressions que la discrimination réalisée à l'aide des 8 paramètres acoustiques sélectionnés dans la première partie de la thèse (v. section 2.3.2.4). Dans l'analyse avec "cross validation", les 7 fonctions dérivées des paramètres acoustiques avaient permis de reclasser 50% des expressions dans leurs catégories d'origine. Les confusions dans la classification des expressions présentent également des caractéristiques différentes, selon que l'analyse discriminante est effectuée sur la base des paramètres acoustiques ou sur la base des caractéristiques vocales perçues. Les "erreurs" de classification dans l'analyse discriminante effectuée à la section 2.3.2.4 étaient dans une large mesure (mais non exclusivement) attribuables à une absence de discrimination entre les émotions de même niveau d'activation. Les confusions présentées dans le tableau 13 sont moins clairement attribuables au niveau d'activation sous-jacent aux émotions exprimées. Les expressions d'anxiété (activation faible), qui sont les moins bien discriminées par l'analyse, sont par exemple "confondues" avec les expressions de joie calme, d'irritation et de tristesse (activation faible), mais également avec les expressions de peur panique et de désespoir (activation forte).

L'influence du niveau d'activation sur les dimensions vocales perçues a été évaluée indirectement en contrôlant la part de variance de ces dimensions expliquée par le niveau d'activation (défini de manière dichotomique – fort versus faible), puis en testant à nouveau les effets de l'émotion exprimée et de la phrase sur les dimensions vocales dans une série d'ANOVAs à mesures répétées.

Tableau 14 : ANOVAs à mesures répétées - Effets "within" de l'émotion sur les jugements moyens obtenus pour les 8 dimensions vocales perçues, avec covariation de la variable 'activation' (2 niveaux).

dimension	source	df	F	sig.	eta^2	dimension	source	df	F	sig.	eta^2
intensité	emo	(7, 56)	12.07	.000	0.60	perçante	emo	(7, 56)	9.13	.000	0.53
hauteur	emo	(7, 56)	4.15	.001	0.34	articulation	emo	(7, 56)	7.76	.000	0.49
intonation	emo	(7, 56)	11.17	.000	0.58	rauque	emo	(7, 56)	3.37	.005	0.30
rapidité	emo	(7, 56)	11.37	.000	0.59	instabilité	emo	(7, 56)	12.66	.000	0.61

Les résultats, présentés dans le tableau 14, en comparaison avec les résultats présentés dans le tableau 11, indiquent que le niveau d'activation représente une composante non-négligeable de l'effet de l'émotion exprimée sur les dimensions vocales perçues. La taille de l'effet de l'émotion diminue sensiblement lorsque l'effet de l'activation est contrôlé. Elle passe pour l'intensité de $\eta^2 = 0.80$ à $\eta^2 = 0.60$, pour la hauteur de $\eta^2 = 0.71$ à $\eta^2 = 0.34$, pour l'intonation de $\eta^2 = 0.80$ à $\eta^2 = 0.58$, pour la qualité perçante de $\eta^2 = 0.82$ à $\eta^2 = 0.53$. La taille des effets pour la rapidité (de $\eta^2 = 0.71$ à $\eta^2 = 0.59$), la qualité rauque (de $\eta^2 = 0.44$ à $\eta^2 = 0.30$), l'instabilité (de $\eta^2 = 0.66$ à $\eta^2 = 0.61$) et, en particulier, pour l'articulation (de $\eta^2 = 0.50$ à $\eta^2 = 0.49$) est affectée dans une moindre mesure par le contrôle du niveau d'activation. Ces différences sont de même importance

que les différences observées pour la taille des effets de l'émotion exprimées sur les paramètres acoustiques avant et après le contrôle du niveau d'activation.

3.4 Discussion

Les corrélations entre les dimensions vocales évaluées soulèvent plusieurs points de discussion qui seront abordés successivement ci-dessous.

Premièrement, on observe dans l'étude principale, une absence de corrélation entre la dimension 'articulation' et la dimension 'rapidité', alors que dans le prétest ces dimensions sont corrélées. Les expressions qui sont jugées plus rapides tendent à être également évaluées (par le même groupe d'auditeurs) comme mieux articulées. La relation observée dans le prétest et l'absence de relation rapportée pour l'étude principale s'opposent à l'idée généralement acceptée qu'un débit de parole plus rapide devrait être lié à une qualité d'articulation dégradée. Sur le plan de la perception, cette relation ne serait donc pas valable pour certains enregistrements de parole émotionnelle. Il n'existe à notre connaissance pas de support théorique ou empirique à l'appui de cette spéculation. Toutefois, on pourrait imaginer qu'un style de "parole traînante" (à la fois lente et mal articulée) pourrait être associée par exemple à la tristesse ou à l'ennui. Les données du prétest présentées dans le tableau 4 indiquent effectivement que les expressions perçues comme plus tristes sont également perçues comme plus lentes ($r = -0.75$) et moins bien articulées ($r = 0.64$). Les graphiques 2 et 6 indiquent d'autre part que les enregistrements qui expriment la tristesse et l'ennui sont perçues comme moins bien articulées et comme plus lentes que les autres expressions. Les graphiques 14 et 16 présentent des résultats identiques sur ce point pour l'étude principale. De plus, les résultats d'un questionnaire construit pour examiner les représentations relatives aux caractéristiques vocales qui correspondraient à différentes émotions (v. annexe B3) indiquent que cette relation entre l'articulation (plus mauvaise) et la rapidité d'élocution (plus lente) existe également au niveau des représentations de la qualité vocale associée à la tristesse (graphiques A1 et A8 de la section B3 en annexe).

En plus de la corrélation entre la rapidité et l'articulation, une autre corrélation, entre la hauteur perçue et la qualité de l'articulation, est significative dans le prétest (tableau 4) et non significative dans l'étude principale (tableau 7). Cette différence pourrait s'expliquer par l'existence d'une relation différente entre ces dimensions dans l'échantillon d'expressions présentées. Nous avons constaté précédemment que la relation entre la rapidité et l'articulation perçues pouvaient s'expliquer en partie par la présence d'expressions de tristesse et d'ennui (les expressions d'intérêt et de dégoût semblent également favoriser cette relation). De même, la présence d'enregistrements exprimant l'intérêt et l'ennui dans le prétest pourrait en partie expliquer la corrélation entre la hauteur et la

qualité de l'articulation. Les expressions d'intérêt sont perçues comme très bien articulées et très aiguës, alors que les expressions d'ennui sont perçues comme très mal articulée et très graves.

Plus généralement, la forte colinéarité entre les dimensions vocales évaluées est au moins en partie liée à la sélection des expressions présentées. De la même manière que les corrélations entre les paramètres acoustiques, présentées dans la première partie de la thèse, étaient amplifiées par la présence d'expressions émotionnelles contrastées sur le plan de l'activation (v. graphique 10 à la section 2.3.2.5), les caractéristiques vocales perçues sont liées par les mécanismes de production vocale qui engendrent des covariations entre certaines dimensions (par exemple entre la hauteur et l'intensité perçue) pouvant être particulièrement importantes pour les expressions sélectionnées. D'un autre côté, la colinéarité pourrait résulter en partie de la difficulté qu'auraient les auditeurs à isoler les dimensions soumises à leur évaluation. Dans cette perspective, les auditeurs formeraient des impressions perceptives plus globales à partir desquelles ils dériveraient leurs jugements pour différentes dimensions vocales en se basant sur des critères de décision similaires.

Indépendamment de sa cause, la colinéarité est en tous cas importante, deux composantes parviennent à rendre compte de près de 75% de la variance observée pour les 8 dimensions vocales évaluées. Cependant une part indépendante de variabilité (même réduite) subsiste pour chacune des dimensions qui parviennent finalement à réaliser une discrimination des émotions exprimées meilleure que la discrimination réalisée à l'aide des paramètres acoustiques.

Cette dernière constatation pourrait également être attribuée à deux facteurs différents. D'une part, des dimensions vocales perçues ont été évaluées, pour lesquelles des corrélats acoustiques, même approximatifs, n'ont pas été mesurés. L'inclusion d'autres mesures acoustiques qui rendraient compte, au moins en partie, de la qualité de l'articulation perçue (par exemple des mesures relatives aux formants), de la stabilité perçue (par exemple une mesure du "jitter") et de la qualité rauque perçue pourraient améliorer la discrimination des émotions exprimées sur le plan acoustique et annuler l'avantage des dimensions vocales perçues sur ce plan. D'autre part, les jugements relatifs aux dimensions vocales perçues pourraient être indirectement affectés par les émotions exprimées qui sont assurément perçues par les auditeurs. La troisième partie de cette thèse qui examine la relation entre les émotions exprimées et les émotions perçue dans les enregistrements utilisés indique que les émotions exprimées par les locuteurs/acteurs sont très aisément reconnues par des groupes d'auditeurs. Si les jugements concernant les dimensions vocales dérivent d'une impression perceptive globale formée par des auditeurs lors de l'écoute des expressions (v. point de discussion ci-dessus) et si la qualité émotionnelle est dominante dans cette impression, elle pourrait influencer les jugements en fonction d'a priori sur la relation entre une émotion (par exemple la 'colère') et des

caractéristiques vocales. Nous avons examiné si de tels a priori (évidemment eux-même dérivé de l'expérience, au sens large, des individus) existent pour les propriétés de la voix émotionnelle à l'aide d'un questionnaire présenté en annexe (B3). Les résultats de ce questionnaire indiquent effectivement qu'il existe des représentations partagées relativement à certaines propriétés de la voix émotionnelle et que ces représentations sont en partie conformes aux jugements obtenus pour les expressions produites par les acteurs. Les données qui ont été présentées n'excluent donc pas la présence d'un effet indirect de l'émotion exprimée (perçue par les auditeurs) sur les jugements relatifs à la qualité vocale. Cet effet pourrait expliquer, en partie, la relation plus étroite entre les caractéristiques vocales perçues et les émotions exprimées, relativement à la relation entre les caractéristiques acoustiques et les émotions exprimées.

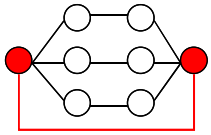
Le fait de soumettre aux jugements des auditeurs des dimensions vocales, conçues comme partiellement indépendantes et destinées à cerner plus globalement la qualité vocale perçue, pose plusieurs problèmes, en partie déjà évoqués dans l'introduction et abordés ci-dessus dans la discussion: on peut douter que les auditeurs isolent spontanément des dimensions de cette sorte lorsqu'ils entendent des expressions vocales; les dimensions sont en partie dépendantes et elles ne capturent pas la totalité de l'impression perceptive. On pourrait en déduire qu'il serait plus judicieux de demander aux auditeurs de décrire librement les caractéristiques des expressions dans le but de dégager des propriétés particulièrement saillantes pour différentes émotions. Malheureusement, nos prétests (v. annexe B1) ont montré que des auditeurs non-spécialistes de la voix ne semblent pas capables de décrire verbalement des caractéristiques vocales qui ne feraient pas référence à des interprétations relatives aux caractéristiques des locuteurs (émotions, attitudes, âge, santé, sexe) qui produisent les expressions décrites. Suivant une proposition de Kreiman & Gerratt (1998), une approche alternative pourrait consister à obtenir des jugements de similarité pour des expressions présentées par paires puis à examiner statistiquement les dimensions sous-jacentes à ces jugements. Ce type d'approche à été utilisé par Green & Cliff (1975) pour l'évaluation d'expressions émotionnelles. Dans cette étude, les expressions ont été différenciées par les auditeurs sur deux axes que les auteurs ont identifiés comme correspondant à la valence et à l'activation sous-jacentes aux expressions émotionnelles présentées. Ces résultats indiquent qu'il n'est probablement pas possible d'identifier des caractéristiques vocales perçues associées aux émotions exprimées à l'aide de cette procédure. Lorsque l'impression perceptive qui se dégage des expressions utilisées est dominée par la qualité émotionnelle, les jugements de similarité seront très probablement basés sur des caractéristiques émotionnelles (telles que la valence ou l'activation) et non sur des propriétés qui décrivent directement la qualité de la voix.

3.5 Conclusion

Les études de jugements présentées ci-dessus démontrent qu'il est possible d'obtenir des jugements cohérents/fidèles relatifs aux propriétés perçues de la voix émotionnelle. De plus, les 8 caractéristiques vocales évaluées parviennent à discriminer les émotions exprimées avec plus de précision que les 8 paramètres acoustiques sélectionnés pour la discrimination au chapitre précédant. Cet "avantage" des caractéristiques vocales perçues sur les paramètres acoustiques mesurés peut être expliqué de différentes manières. Premièrement, l'évaluation des caractéristiques vocales perçues permet de saisir des aspects vocaux qui ne sont pas capturés, ou qui sont difficiles à mesurer, au niveau acoustique. Dans le cas présent, les caractéristiques vocales perçues "rauque", "instabilité" et "qualité de l'articulation" contribuent à la discrimination des émotions exprimées et sont comparativement peu corrélées aux paramètres acoustiques utilisés pour la discrimination au chapitre précédent. Deuxièmement, il est possible que l'impression perceptive formée par les auditeurs soit dominée par la qualité émotionnelle des expressions et que les jugements relatifs à la qualité vocale soient biaisés dans le sens d'une conformation des caractéristiques vocales évaluées à des représentations/stéréotypes associés à la qualité émotionnelle perçue. Cette deuxième possibilité pourrait être explorée, d'une part, en développant une meilleure connaissance des représentations (ou stéréotypes) associées aux expressions vocales émotionnelles afin d'évaluer dans quelle mesure ces représentations sont susceptibles de biaiser les jugements relatifs à la qualité vocale perçue et, d'autre part, en comparant les jugements relatifs à la qualité vocale obtenus pour des expressions émotionnelles avec des jugements relatifs à la qualité vocale obtenus pour des expressions non-émotionnelles. Dans la mesure où la relation entre les propriétés acoustiques des expressions et les jugements relatifs à la qualité vocale serait très semblable pour ces deux types d'expressions (émotionnelles versus non émotionnelles), la possibilité d'un biais dans les jugements introduit par l'impression émotionnelle pourrait être écartée.

Par ailleurs, il est apparu que les évaluations des caractéristiques vocales perçues présentent un potentiel heuristique certain. Dans l'étude présentée, elles ont notamment permis de formuler des hypothèses relativement à des associations inhabituelles de caractéristiques vocales qui seraient propres à la voix émotionnelle. Cette approche est donc susceptible de générer de nouvelles hypothèses relativement aux aspects vocaux qui sont spécifiques à la voix émotionnelle et qui permettraient de la caractériser.

4 Emotions attribuées aux expressions vocales



4.1 Introduction

Dans les chapitres précédents, le processus de communication vocale des émotions a été décomposé en étapes successives conformément au 'lens model' de Brunswik proposé par Scherer (1978, 2003) pour l'étude de la communication non-verbale. Les caractéristiques acoustiques d'un ensemble d'expressions vocales émotionnelles ont d'abord été analysées, puis les caractéristiques vocales perçues de ces mêmes expressions ont été étudiées. Dans une troisième étape présentée ci-dessous, les attributions émotionnelles relatives à ces expressions vocales seront examinées.

4.1.1 Evaluer les attributions émotionnelles

La plupart des auditeurs exposés à des enregistrements d'expressions vocales effectuent probablement des inférences spontanées relativement à différentes caractéristiques propres aux locuteurs enregistrés. On sait notamment que des caractéristiques telles que l'âge, le sexe, les émotions, les attitudes ou même la personnalité des locuteurs peuvent être inférées à partir d'enregistrements vocaux relativement brefs. Les méthodes utilisées pour étudier les différentes attributions effectuées par les auditeurs sont en revanche très variables et représentent une problématique centrale dans ce domaine d'étude.

Les attributions émotionnelles peuvent, par exemple, être obtenues en demandant aux auditeurs de décrire librement (en utilisant leurs propres termes) les émotions communiquées par les locuteurs. Une autre possibilité consiste à soumettre plusieurs réponses alternatives aux auditeurs qui auront alors pour tâche de choisir la réponse la plus appropriée (ou les réponses les plus appropriées) relativement à l'émotion (ou aux émotions) qu'ils perçoivent. Cette deuxième possibilité a été en général privilégiée par les auteurs des recherches effectuées dans ce domaine. L'objectif des études effectuées dans le passé étant généralement d'évaluer dans quelles mesures différentes émotions exprimées peuvent être correctement identifiées par un groupe d'auditeurs, cette méthode semble à première vue la plus adaptée. Les alternatives de réponse proposées aux auditeurs sont choisies de manière à correspondre à la gamme des émotions théoriquement exprimées dans les expressions présentées et la proportion d'auditeurs qui sélectionnent l'alternative correcte pour chaque expression est utilisée comme indicateur du pouvoir de communication de l'expression. Comparativement, les réponses libres présentent l'inconvénient de devoir être (re)catégorisées par les chercheurs avant de pouvoir effectuer une estimation de la proportion des réponses correctes obtenues pour chaque expression. Cependant, la méthode qui consiste à demander aux auditeurs de

sélectionner la meilleure réponse parmi un nombre limité d'alternatives comporte également des désavantages. Les attributions effectuées ne découlent notamment pas d'une véritable reconnaissance des émotions exprimées mais plutôt d'une discrimination entre les alternatives proposées. Cette distinction entre reconnaissance et discrimination est particulièrement cruciale dans les situations – malheureusement fréquentes dans les études effectuées dans le passé – où une seule alternative de réponse désigne une émotion positive (par exemple la joie) ou encore lorsqu'une seule alternative de réponse désigne une émotion faiblement activée (par exemple la tristesse). Dans ces cas, la tâche des auditeurs est facilitée, au sens où il est possible d'identifier la réponse correcte en ne reconnaissant pas réellement la catégorie émotionnelle (joie ou tristesse) mais en identifiant une émotion positive (ou l'absence d'émotion négative) ou une émotion faiblement activée (ou l'absence d'une activation forte).

Une autre approche consiste à demander aux auditeurs d'effectuer pour chaque expression présentée des jugements relatifs à des dimensions continues (unipolaires ou bipolaires) telles que la valence, l'activation ou l'intensité des émotions exprimées. Dans l'étude présentée ci-dessous, une variante de cette approche a été adoptée. Les émotions exprimées par les acteurs dans cette étude sont la colère froide et la colère chaude, la joie calme et la joie intense, l'anxiété et la peur panique, la tristesse et le désespoir. Ces émotions ont été théoriquement définies comme correspondant à 4 familles émotionnelles (colère, joie, peur et tristesse) croisées avec deux niveaux d'activation émotionnelle (activation faible et activation forte). En conséquence, les auditeurs ont été priés d'indiquer pour chaque expression l'intensité de la colère perçue, l'intensité de la joie perçue, l'intensité de la peur perçue et l'intensité de la tristesse perçue. Cette méthode présente l'avantage de supprimer (ou au moins d'atténuer) le problème de la discrimination et permet d'obtenir des attributions sur des échelles continues qui peuvent ensuite être analysées sur le plan statistique en utilisant des méthodes paramétriques.

Les jugements d'intensité émotionnelle comportent toutefois également leurs inconvénients. Premièrement, on ne peut exclure une confusion partielle entre l'évaluation de l'intensité émotionnelle perçue et le degré de (in)certitude des jugements formulés par les auditeurs. Dans cette optique, des jugements élevés d'intensité émotionnelle perçue pourraient refléter en partie une certitude importante concernant la présence d'un certain type d'émotion, alors que des jugements d'intensité faibles refléteraient une forme d'incertitude relativement à la présence d'une émotion donnée. A notre avis, cette possible confusion sur le plan de la nature des jugements obtenus peut toutefois être contrôlée en donnant des instructions appropriées aux auditeurs. De plus, le degré de certitude relativement à la présence/absence d'un certain type d'émotion pourrait être en partie lié à

l'intensité émotionnelle perçue au sens où des émotions très intenses seraient effectivement plus clairement identifiables que des émotions peu intenses.

D'autre part, dans le cadre de l'étude présentée ci-dessous, ce n'est pas l'intensité émotionnelle mais l'activation émotionnelle qui a été manipulée sur le plan expressif. Or l'intensité et l'activation correspondent à des aspects de la réaction émotionnelle qui sont également partiellement confondus. L'intensité émotionnelle se rapporte au sentiment associé à la réaction émotionnelle alors que l'activation émotionnelle se rapporte plus directement à la composante physiologique (et par extension aux expressions vocales) de la réaction émotionnelle. Ainsi, une expression qui correspond à un état émotionnel avec une composante d'activation forte sera souvent, mais pas nécessairement, associé à un sentiment émotionnel intense.

Dans l'étude présentée ci-dessous, les instructions données aux auditeurs ont été formulées de manière à obtenir des évaluations ayant trait spécifiquement à l'intensité émotionnelle et non à l'activation émotionnelle ou à la (in)certitude concernant la présence d'un certain type d'émotion. Les méthodes utilisées et les résultats obtenus sont décrits ci-dessous.

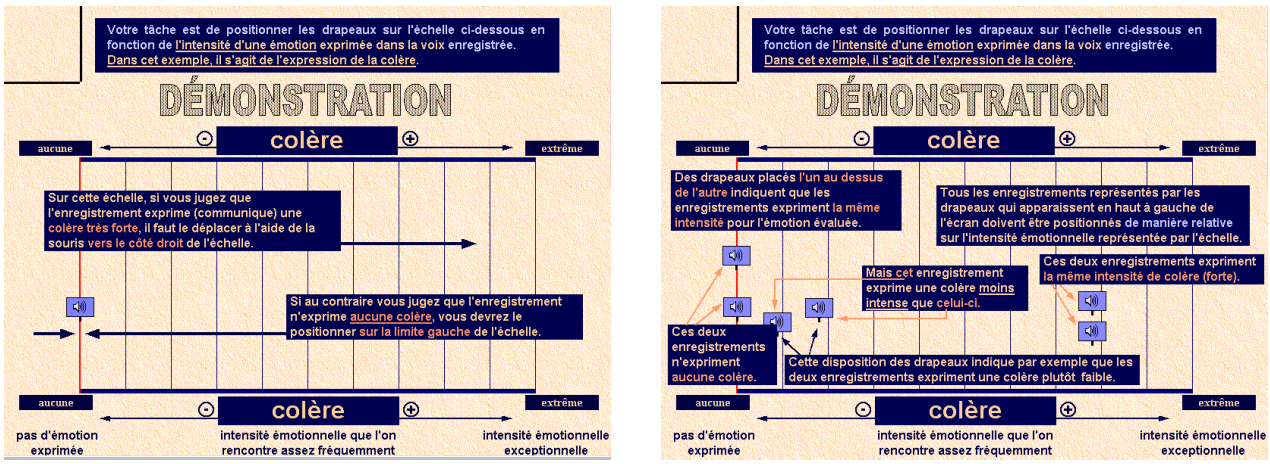
4.2 Méthode

La méthode utilisée pour accéder aux attributions émotionnelles a été développée spécifiquement pour les besoins de cette étude. Elle résulte de la décision d'obtenir des jugements relatifs à l'intensité émotionnelle perçue pour plusieurs types d'émotions et s'inspire de la méthode utilisée au chapitre précédent pour obtenir les jugements relatifs aux propriétés vocales perçues.

Au lieu de présenter une expression et de demander aux auditeurs d'évaluer simultanément l'intensité émotionnelle pour plusieurs types d'émotions, une échelle d'intensité émotionnelle pour un type d'émotion donné est présentée aux auditeurs sur un écran d'ordinateur. Les expressions produites par un même locuteur sont successivement présentées sur le même écran sous forme d'icônes identiques. L'auditeur peut écouter les expressions aussi souvent qu'il le souhaite en double-cliquant sur ces icônes. Sa tâche consiste à placer les expressions de manière relative sur l'échelle qui représente l'intensité émotionnelle ; une position à l'extrême gauche de l'échelle indique une intensité émotionnelle nulle (absence du type d'émotion considéré) ; une position à l'extrême droite de l'échelle indique une intensité émotionnelle extrême. Les réponses des auditeurs sont transcrites sur une échelle continue de 0 (extrême gauche de l'échelle) à 10 (extrême droite de l'échelle). La figure 1 représente une partie des instructions séquentiellement présentées aux auditeurs au début de l'étude. Un soin particulier a été apporté à la définition de l'*intensité* émotionnelle en présentant un exemple animé illustrant la signification de différentes positions sur l'échelle au début de l'étude. L'objectif de ces instructions détaillées était de s'assurer que les

auditeurs comprennent qu'il s'agit d'indiquer l'*intensité* émotionnelle perçue et non une autre dimension associée telle que l'activation ou la (in)certitude concernant la présence d'un certain type d'émotion.

Figure 1: Ecrans d'instructions présentés au début de l'étude (un exemple animé illustre séquentiellement la signification de différentes positions sur l'échelle)



Dans une situation plus classique où plusieurs échelles émotionnelles sont présentées directement après l'écoute d'une expression, certains auditeurs ont tendance à utiliser une stratégie de discrimination en sélectionnant une intensité assez forte sur une échelle et en indiquant une intensité nulle sur les autres échelles. Un avantage de la procédure décrite ci-dessus consisterait donc à réduire sensiblement la quantité des réponses nulles qui ont tendance à apparaître lorsque les échelles émotionnelles sont présentées simultanément. Un autre avantage est de focaliser l'attention des auditeurs sur un seul type d'attribution afin d'éviter que la nature des attributions effectuées ne se modifie au cours de la session. Plus généralement, la possibilité de comparer directement toutes les expressions produites par un même locuteur permet à l'auditeur de répondre relativement à la gamme des expressions produites par ce locuteur. De cette manière, lorsqu'un auditeur indique une intensité émotionnelle très forte pour une expression et se trouve par la suite confronté à une autre expression qu'il juge encore plus intense, il dispose de la possibilité de "corriger" sa réponse précédente.

4.2.1 Expressions utilisées et dimensions émotionnelles évaluées

Des attributions relatives à l'intensité de joie perçue, à l'intensité de peur perçue, à l'intensité de colère perçue et à l'intensité de tristesse perçue ont été obtenues pour les 144 expressions émotionnelles décrites dans les chapitres précédents. Pour rappel, ces expressions ont été produites par 9 acteurs qui expriment 8 émotions différentes en prononçant deux séquences de syllabes sans signification. Les séquences de syllabes sont: 1. hätt sandik prong nju ventsie, 2. fi gött leich jean kill gos terr. Les émotions exprimées par les acteurs sont la colère froide ('irrit') et la colère chaude

('rage'), la joie calme ('joie') et la joie intense ('exalt'), l'anxiété ('anx') et la peur panique ('paniq'), la tristesse ('trist') et le désespoir ('desp').

4.2.2 Auditeurs et procédure

Les 144 expressions émotionnelles ont été divisées en 4 groupes d'expressions qui ont été évalués par 4 groupes indépendants d'auditeurs. Ces groupes ont été créés afin de ne pas surcharger les ressources attentionnelles des auditeurs en leur demandant d'évaluer un nombre trop important d'expressions durant une même session. Chacun des groupes d'auditeurs était composé de 14 étudiants de 1ère année en Psychologie à l'Université de Genève. Ces étudiants ont participé à cette étude en échange d'un crédit de cours; ils étaient tous francophones et ne souffraient d'aucun déficit auditif diagnostiqué. Chaque groupe a évalué les expressions produites par 3 locuteurs différents, soit au total 48 expressions émotionnelles (3 locuteurs*2 phrases*8 émotions exprimées). Afin de tester l'absence de différences systématiques entre les groupes, les enregistrements produits par le locuteur 3 ont été évalués par les 4 groupes.

Un premier groupe constitué de 10 femmes et 4 hommes (âge moyen = 21.5 ans, sd = 4.0) a évalué les expressions produites par l'actrice 2, l'acteur 3 et l'acteur 4 ; un deuxième groupe composé de 11 femmes et 3 hommes (âge moyen = 22.3, sd = 5.7) a jugé les expressions produites par l'actrice 7, l'acteur 8 et l'acteur 3 ; un troisième groupe composé de 12 femmes et 2 hommes (âge moyen = 21.5, sd = 3.0) a évalué les expressions produites par l'actrice 9, l'actrice 10 et l'acteur 3 ; et un quatrième groupe composé de 12 femmes et 2 hommes (âge moyen = 23.3, sd = 7.6) a évalué les expressions produites par l'actrice 11, l'acteur 12 et l'acteur 3.

Pour chaque auditeur, un ordre aléatoire de présentation des expressions est défini par l'ordinateur au début de chaque session. Un locuteur est sélectionné aléatoirement parmi les 3 locuteurs que chaque auditeur est chargé d'évaluer. Les expressions du premier locuteur sélectionné sont évaluées successivement sur les quatre échelles d'intensité émotionnelle (l'intensité de joie, de peur, de colère et de tristesse perçues). L'ordre de présentation de ces échelles est défini aléatoirement par l'ordinateur. Un deuxième locuteur est ensuite choisi aléatoirement parmi les 2 locuteurs restant et un nouvel ordre de présentation aléatoire des 4 échelles est défini. Finalement, un troisième ordre aléatoire de présentation des échelles d'intensité émotionnelle est défini pour l'évaluation des expressions produites par le dernier locuteur. D'autre part, pour chaque échelle présentée, l'ordinateur définit un nouvel ordre aléatoire de présentation des expressions produites par le locuteur évalué.

4.3 Résultats

4.3.1 Evaluation des différences de réponse entre les groupes d'auditeurs

Les expressions produites par le locuteur 3 (16 expressions) ont été évaluées par l'ensemble des auditeurs afin de permettre de vérifier, pour ces expressions, que les jugements produits par les auditeurs ne sont pas dépendants des 4 groupes d'auditeurs créés. Quatre ANOVAs à mesures répétées (avec un facteur 'expression' intra/within à 16 niveaux et un facteur 'groupe' inter/between à 4 niveaux) ont été effectuées pour tester l'existence d'un éventuel effet (inter/between) du groupe d'auditeurs sur les jugements effectués pour chacune des 4 intensités émotionnelles évaluées. L'effet principal du groupe est non-significatif pour les 4 échelles. Les résultats de ces ANOVAs sont représentés dans le tableau 1.

Tableau 1: Effets inter/between du groupe d'auditeurs sur les 4 jugements d'intensité émotionnelle pour les expressions produites par le locuteur 3

	df	F	sig.	eta^2
joie	(3, 52)	1.87	.146	0.10
peur	(3, 52)	1.83	.153	0.10
tristesse	(3, 52)	0.63	.599	0.04
colère	(3, 52)	0.27	.848	0.02

4.3.2 Fidélité inter-auditeurs des jugements

Afin d'évaluer la fidélité inter-auditeurs des jugements obtenus, des coefficients de corrélations intraclasses ont été calculés pour les 4 types de jugements effectués par chaque groupe d'auditeurs. Ces coefficients sont représentés dans le tableau 2. Dans ce tableau, le coefficient r correspond à une estimation de la corrélation moyenne entre les jugements de toutes les paires d'auditeurs et le coefficient R représente l'indice de fidélité des jugements (équivalent à la valeur α de Cronbach). Ces valeurs sont calculées sur la base de 48 jugements pour chaque groupe d'auditeur et chaque échelle émotionnelle; les 4 groupes sont composés de 14 auditeurs.

Tableau 2: Coefficients de fidélité (r = single measure intraclass correlation, R = average measure intraclass correlation) pour les 4 jugements d'intensité émotionnelle et les 4 groupes d'auditeurs.

intensité	groupe 1 (N=14)		groupe 2 (N=14)		groupe 3 (N=14)		groupe 4 (N=14)	
émotionnelle	r	R	r	R	r	R	r	R
joie	0.55	0.95	0.47	0.93	0.48	0.93	0.41	0.91
peur	0.54	0.94	0.49	0.93	0.63	0.96	0.44	0.92
tristesse	0.65	0.96	0.53	0.94	0.65	0.96	0.66	0.96
colère	0.66	0.96	0.64	0.96	0.67	0.97	0.70	0.97

Tous les indices de fidélité (R) sont compris entre 0.91 et 0.97; la fidélité inter-auditeurs est donc très satisfaisante. Les corrélations entre les jugements effectués par les auditeurs tendent toutefois à

être plus élevées pour les jugements relatifs à l'intensité perçue de colère et de tristesse que pour les jugements relatifs à l'intensité perçue de peur et de joie.

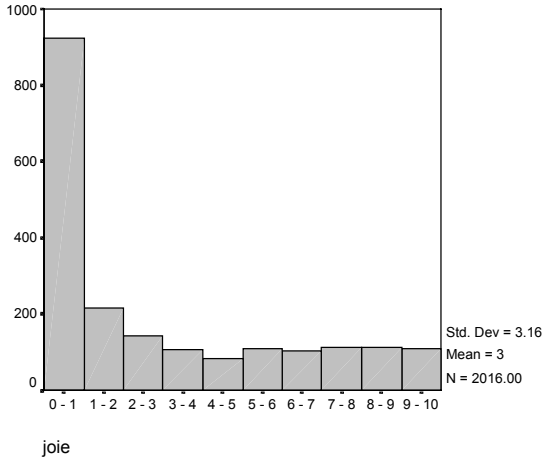
4.3.3 Distributions des réponses

La fidélité inter-auditeurs des jugements étant établie, des jugements moyens peuvent être calculés pour chacune des 144 expressions émotionnelles et pour chacune des 4 intensités émotionnelles évaluées. Au préalable, un ajustement a été effectué relativement aux jugements obtenus pour les expressions produites par le locuteur 3. Les expressions produites par ce locuteur ont été évaluées par 4 fois plus d'auditeurs que les autres expressions. Dès lors, des moyennes calculées pour les expressions produites par ce locuteur seraient basées sur 4 fois plus de jugements que les moyennes obtenues pour les 8 autres locuteurs. En conséquence, une partie des jugements effectués pour le locuteur 3 ont été supprimés aléatoirement. Les jugements de 4 auditeurs du 1er groupe, de 4 auditeurs du 2ème groupe, de 3 auditeurs du 3ème groupe et de 3 auditeurs du 4ème groupe d'auditeurs ont été sélectionnés aléatoirement; les jugements effectués par les autres auditeurs pour les expressions produites par le locuteur 3 ont été éliminés du calcul de la moyenne. Toutes les moyennes sont ainsi basées sur 14 jugements pour chaque expression.

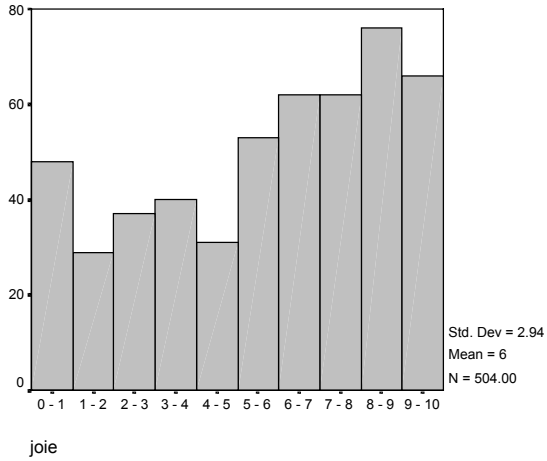
Avant d'examiner les jugements moyens pour chaque expression, les distributions des jugements bruts ont été examinées. Les histogrammes 1 à 8 représentent les distributions des jugements de joie (graphiques 1 et 2), de peur (graphiques 3 et 4), de tristesse (graphiques 5 et 6) et de colère (graphiques 7 et 8). Les graphiques présentés à gauche (graphiques 1, 3, 5 et 7) représentent les jugements obtenus pour l'ensemble des expressions (144 expressions * 14 jugements). A droite, le graphique 2 représente les jugements de joie obtenus pour les expressions de joie calme et de joie intense (36 expressions * 14 jugements), le graphique 4 représente les jugements de peur obtenus pour les expressions d'anxiété et de peur panique (36 expressions * 14 jugements), le graphique 6 représente les jugements de tristesse obtenus pour les expressions de tristesse et de désespoir (36 expressions * 14 jugements) et le graphique 8 représente les jugements de colère obtenus pour les expressions de colère froide et de colère chaude (36 expressions * 14 jugements).

Les distributions des jugements pour la totalité des expressions (graphiques 1, 3, 5 et 7) indiquent que les auditeurs ont attribué une intensité quasi nulle à une quantité importante d'expressions pour les 4 échelles d'intensité émotionnelle. La proportion des jugements d'intensité compris entre 0 et 1 se situe entre 40% et 50% pour les 4 échelles.

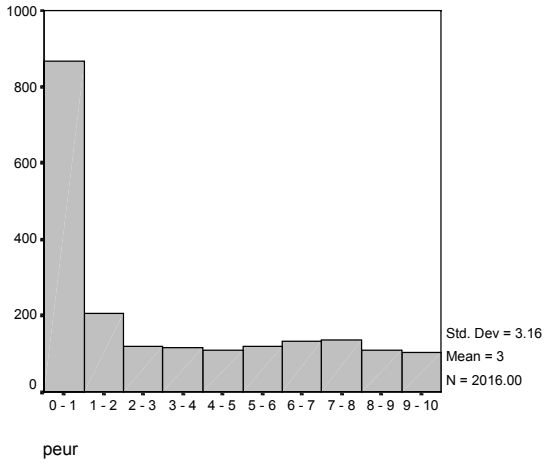
Graphique 1: Intensité de joie perçue, distribution des jugements pour la totalité des expressions



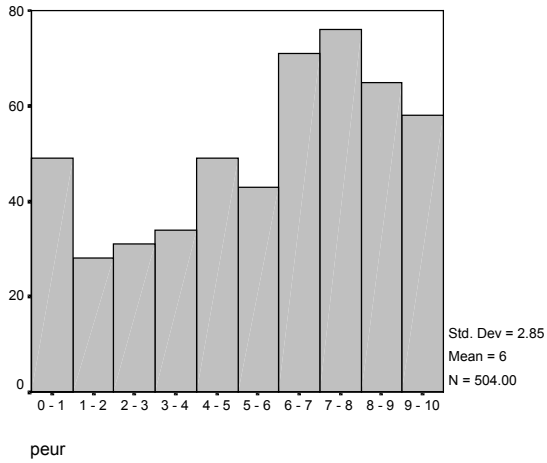
Graphique 2: Intensité de joie perçue, distribution des jugements pour les expressions de joie calme et de joie intense



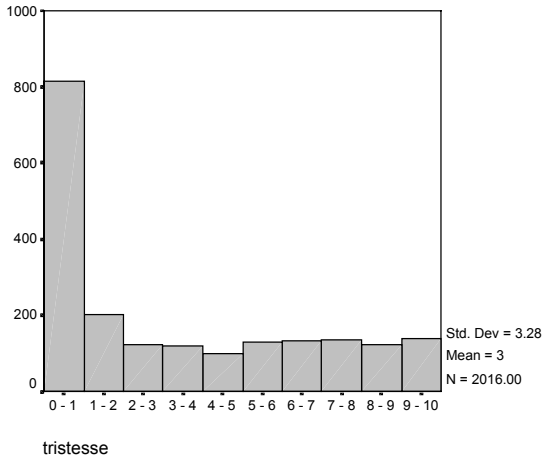
Graphique 3: Intensité de peur perçue, distribution des jugements pour la totalité des expressions



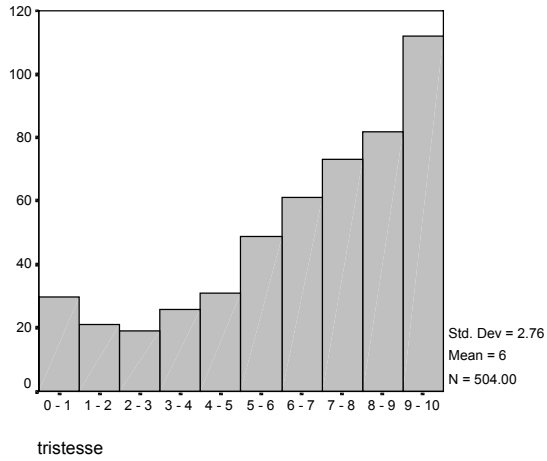
Graphique 4: Intensité de peur perçue, distribution des jugements pour les expressions d'anxiété et de peur panique



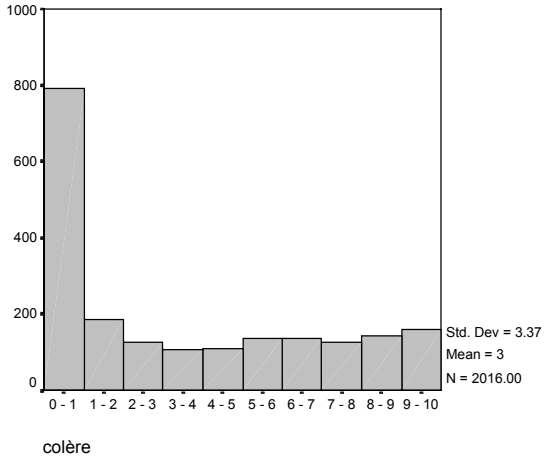
Graphique 5: Intensité de tristesse perçue, distribution des jugements pour la totalité des expressions



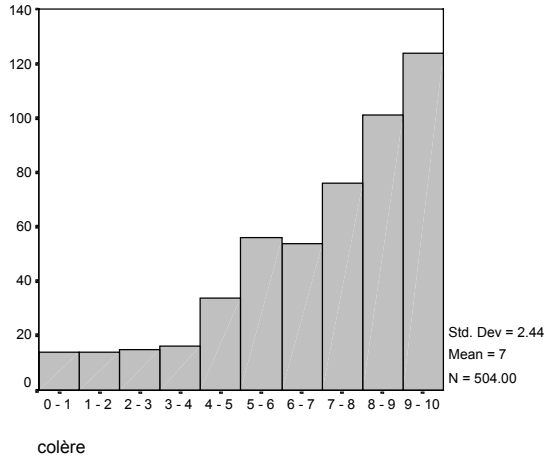
Graphique 6: Intensité de tristesse perçue, distribution des jugements pour les expressions de tristesse et de désespoir



Graphique 7: Intensité de colère perçue, distribution des jugements pour la totalité des expressions



Graphique 8: Intensité de colère perçue, distribution des jugements pour les expressions de colère froide et de colère chaude



Théoriquement, seules 25% des expressions sont supposées communiquer une intensité relativement importante sur chaque échelle émotionnelle (les distributions des jugements pour ces expressions sont représentées sur les graphiques 2, 4, 6 et 8). Pour chaque échelle émotionnelle, 75% des expressions sont donc supposées communiquer une intensité faible ou nulle. En conséquence, les distributions des réponses pour la totalité des expressions reflètent d'une part la nature des expressions évaluées (dans chaque cas, un grand nombre d'expressions ne communiquent pas le type d'émotion considéré); d'autre part, la forte proportion de réponses comprises entre 0 et 1 pourrait refléter également une tendance à effectuer une décision binaire en matière de jugement émotionnel, c'est-à-dire à identifier la *présence* versus l'*absence* de l'émotion considérée indépendamment de la modalité de jugement proposée (ici une modalité de jugement continue).

D'autres aspects des distributions semblent toutefois indiquer la présence de jugements continus. Dans les graphiques 1, 3, 5 et 7 les réponses comprises entre 1 et 10 sont réparties très uniformément (toutes les valeurs sont utilisées de manière relativement égale). Les jugements relatifs à l'intensité de joie perçue pour les expressions de joie calme et de joie intense (graphique 2), ainsi que les jugements de peur perçue pour les expressions d'anxiété et de peur panique (graphique 4) présentent un léger pic aux niveaux des jugements compris entre 0 et 1 mais sont, par ailleurs, distribués sur la totalité des deux échelles avec un nombre plus important de jugements d'intensité élevés que de jugements d'intensité faibles. Relativement à l'aspect continu des jugements, les jugements de tristesse pour les expressions de tristesse et de désespoir (graphique 6) et les jugements de colère pour les expressions de colère froide et chaude (graphique 8) présentent des distributions idéales avec un accroissement progressif de la fréquence des jugements obtenus en fonction de l'intensité émotionnelle perçue.

4.3.4 Relations entre les émotions perçues

Quatre jugements d'intensité moyens – une moyenne pour l'intensité perçue de peur, une moyenne pour l'intensité de colère, une moyenne pour l'intensité de tristesse et une moyenne pour l'intensité de joie – basés sur les 14 jugements obtenus (ou retenus dans le cas des expressions produites par le locuteur 3) ont été calculés pour chacune des 144 expressions émotionnelles.

Les corrélations entre ces 4 jugements d'intensité émotionnelle révèlent que les 4 types d'intensité émotionnelle perçue ne sont pas indépendants. L'intensité de peur est corrélée positivement à l'intensité de tristesse. L'intensité de joie est corrélée négativement à l'intensité de tristesse et de colère. L'intensité de tristesse est également corrélée négativement avec l'intensité de colère. Les valeurs des corrélations entre les 4 échelles d'intensité émotionnelle perçue sont représentées dans le tableau 3. Ces corrélations sont toutefois relativement faibles, la variance partagée la plus importante – entre l'intensité de peur perçue et l'intensité de tristesse perçue – n'est que de 20%.

Tableau 3: Corrélations entre l'intensité émotionnelle perçue de joie, de peur, de tristesse et de colère (N = 144), les corrélations significatives ($p < .05$) sont indiquées en gras.

	intensité de joie	intensité de peur	intensité de tristesse
intensité de peur	-0.11		
intensité de tristesse	**-0.35**	**0.45**	
intensité de colère	**-0.27**	-0.04	**-0.35**

La corrélation positive entre l'intensité de tristesse perçue et l'intensité de peur perçue suggère la possibilité de l'existence d'une confusion entre la tristesse et la peur perçue, au sens où les expressions qui reçoivent des jugements élevés de peur reçoivent également des jugements élevés de tristesse.

4.3.5 Relations entre émotions perçues et exprimées, effets des phrases et des locuteurs

L'effet des émotions exprimées (8 types) et des 2 phrases prononcées sur les jugements d'intensité émotionnelle a été évalué par 4 ANOVAs à mesures répétées. Les résultats de ces ANOVAs sont présentés dans le tableau 4. Ces résultats indiquent, de manière prévisible que les 4 jugements d'intensité émotionnelle dépendent très fortement du type d'émotion exprimée. Un effet de la phrase est également apparu, mais uniquement pour les jugements de tristesse. Cet effet correspond à un jugement moyen de l'intensité de tristesse égal à 3,5 (écart-type: 2.4) pour la première phrase et égal à 3 (écart-type: 2.3) pour la deuxième phrase. Cette différence d'un demi-point entre les attributions réalisée pour la première et la deuxième phrase représente une différence relativement faible au regard des écart-types qui sont assez importants.

Tableau 4: ANOVAs à mesures répétées - Effets "within" de la phrase et de l'émotion sur les jugements moyens obtenus pour les 4 intensités émotionnelles perçues, les résultats significatifs (p<.05) sont indiqués en gras.

émotion perçue	source	df	F	sig.	eta^2
joie	phrase	(1, 8)	1.62	.239	0.17
	emo	**(7, 56)**	**42.92**	**.000**	**0.84**
	phrase*emo	(7, 56)	1.95	.078	0.20
peur	phrase	(1, 8)	0.61	.459	0.07
	emo	**(7, 56)**	**46.22**	**.000**	**0.85**
	phrase*emo	(7, 56)	0.53	.810	0.06
tristesse	**phrase**	**(1, 8)**	**10.68**	**.011**	**0.57**
	emo	**(7, 56)**	**42.29**	**.000**	**0.84**
	phrase*emo	(7, 56)	0.56	.784	0.07
colère	phrase	(1, 8)	0.85	.382	0.10
	emo	**(7, 56)**	**65.47**	**.000**	**0.89**
	phrase*emo	(7, 56)	1.82	.102	0.18

Tableau 5: ANOVAs - Effets *random* du locuteur et effets d'interactions locuteur*émotion exprimée et locuteur*phrase sur les jugements moyens obtenus pour les 4 intensités émotionnelles perçues, les résultats significatifs (p<.05) sont indiqués en gras.

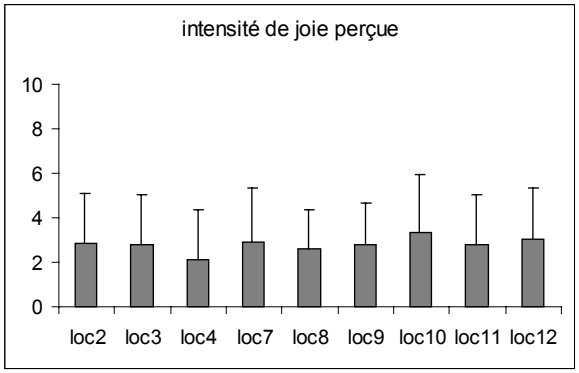
émotion perçue	source	df	F	sig.	eta^2
joie	locuteur	(8, 20)	1.15	.372	0.31
	emo * locuteur	**(56, 56)**	**2.08**	**.004**	**0.68**
	phrase * locuteur	(8, 56)	0.72	.675	0.09
peur	**locuteur**	**(8, 33)**	**2.36**	**.039**	**0.36**
	emo * locuteur	**(56, 56)**	**4.03**	**.000**	**0.80**
	phrase * locuteur	(8, 56)	1.71	.116	0.20
tristesse	locuteur	(8, 33)	1.55	.179	0.27
	emo * locuteur	**(56, 56)**	**3.57**	**.000**	**0.78**
	phrase * locuteur	(8, 56)	1.30	.264	0.16
colère	**locuteur**	**(8, 19)**	**3.46**	**.012**	**0.59**
	emo * locuteur	**(56, 56)**	**2.14**	**.003**	**0.68**
	phrase * locuteur	(8, 56)	1.26	.285	0.15

D'autre part, un effet inter/between très important a été constaté dans les 4 ANOVAs à mesures répétées. Cette observation présuppose l'existence de différences attribuables aux locuteurs ou à des interactions entre les locuteurs et les émotions exprimées. Afin d'évaluer l'importance respective de ces deux types d'effets, une nouvelle ANOVA a été effectuée pour chaque échelle d'intensité émotionnelle avec 2 facteurs *fixes* 'émotion exprimée' (8 niveaux) et 'phrase' (2 niveaux) et un facteur *random* 'locuteur' (9 niveaux). En ce qui concerne les effets de phrases et des émotions exprimées, les résultats de ces ANOVAs sont identiques aux résultats présentés dans le tableau 4. Les résultats obtenus pour l'effet du locuteur, pour l'effet d'interaction entre le locuteur et l'émotion exprimée et pour l'effet d'interaction entre le locuteur et la phrase sont présentés dans le tableau 5.

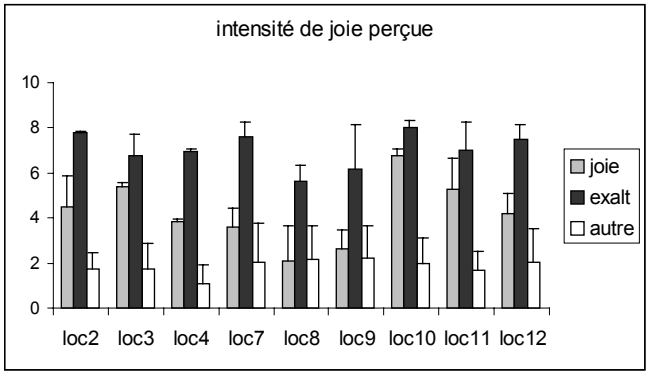
L'effet principal du locuteur est significatif pour l'intensité de peur perçue et pour l'intensité de colère perçue. Les effets de l'interaction entre l'émotion exprimée et le locuteur sont plus importants et sont significatifs pour les 4 types d'intensité émotionnelle évalués.

Les graphiques 9 à 12 illustrent ces effets. Les graphiques 9 et 11 représentent l'effet du locuteur sur l'intensité de joie perçue (graphique 9, l'effet n'est pas significatif) et sur l'intensité de colère perçue (graphique 11, l'effet est significatif). Dans ces graphiques, les moyennes et les écarts-types sont basés sur 16 observations. On observe que les différences entre les jugements moyens de colère obtenus pour différents locuteurs sont relativement faibles. Ils varient de 2.5 (pour le locuteur 4) à 4.5 (pour le locuteur 9). Le graphique 10 représente l'effet du locuteur et du type d'émotion exprimée sur l'intensité perçue de joie pour les expressions de joie calme, pour les expressions de joie intense et pour toutes les autres expressions confondues. Le graphique 12 représente l'effet du locuteur et du type d'émotion exprimée sur l'intensité perçue de colère pour les expressions de colère froide, pour les expressions de colère chaude et pour toutes les autres expressions confondues. Dans ces graphiques, les moyennes et les écarts-types sont basés sur 2 observations lorsque les barres correspondent à un seul type d'émotion exprimée et sur 12 observations lorsque les barres regroupent toutes les autres expressions. Les valeurs représentées sur le graphique 12 suggèrent que les différences sur le plan de l'intensité de la colère perçue pour différents locuteurs pourraient être en partie liées à une moins bonne compétence de certains locuteurs à communiquer volontairement la colère. On observe toutefois aussi nettement que certains locuteurs dans cet échantillon (e.g. le locuteur 9) sont perçus comme exprimant plus de colère que d'autres locuteurs (e.g. le locuteur 4), indépendamment du type d'émotion qu'ils essaient de communiquer.

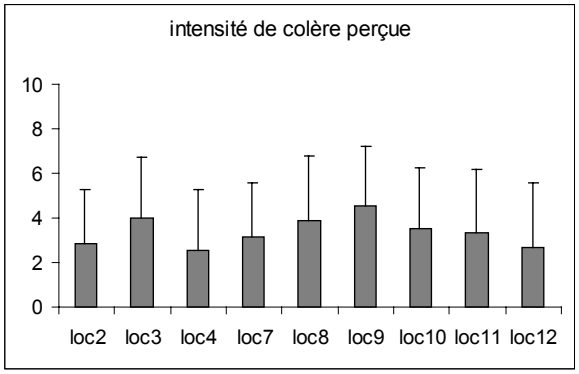
Graphique 9: Moyennes et écarts-types de l'intensité de joie perçue en fonction du locuteur (N=16)



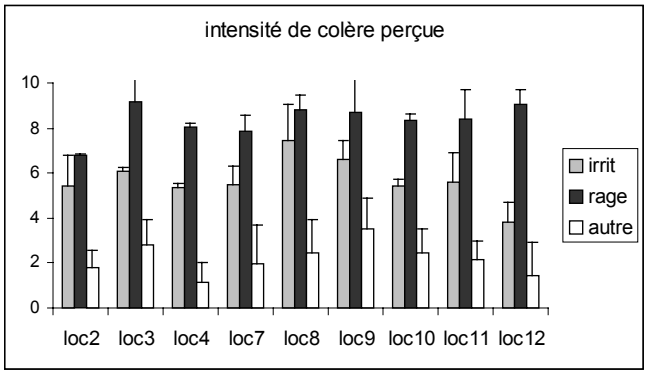
Graphique 10: Moyennes et écarts-types de l'intensité de joie perçue en fonction du locuteur pour les expressions de joie calme et intense (N=2) et pour les autres expressions (N=12)



Graphique 11: Moyennes et écarts-types de l'intensité de colère perçue en fonction du locuteur (N=16)



Graphique 12: Moyennes et écarts-types de l'intensité de colère perçue en fonction du locuteur pour les expressions de colère froide et chaude (N=2) et pour les autres expressions (N=12)



La mise en perspective de l'intensité de colère perçue et de l'intensité de joie perçue fait également apparaître une autre forme d'interaction : les expressions de joie calme des locuteurs 8 et 9 reçoivent des jugements d'intensité de joie qui tendent à être plus faibles que pour les expressions de joie calme produites par les autres locuteurs. En revanche, les expressions de colère froide des locuteurs 8 et 9 reçoivent des jugements d'intensité de colère qui tendent à être plus élevés que pour les expressions de colère froide produites par les autres locuteurs. Cet exemple illustre le fait qu'un même locuteur peut échouer à communiquer clairement un certain type d'émotion, alors qu'il réussira à communiquer un autre type d'émotion.

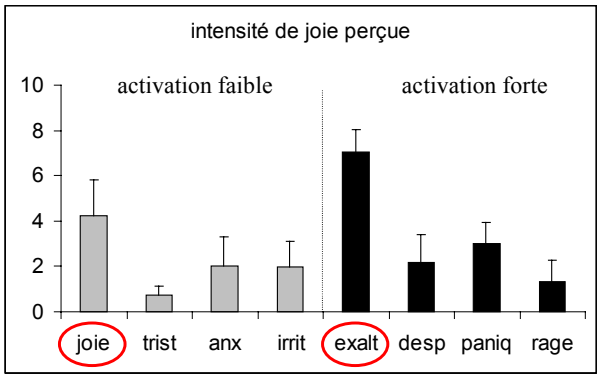
Le tableau 5 et les graphiques 13 à 16 représentent les jugements moyens et les écarts-types par émotion exprimée pour l'intensité perçue de joie, de peur, de tristesse et de colère. Ces valeurs sont basées sur 18 observations (9 locuteurs * 2 phrases). Les résultats de 4 tests post-hoc de Tukey (HSD pour des ANOVAs avec un facteur 'émotion exprimée') sont présentés sous les graphiques, ils signalent les différences significatives entre les moyennes pour les différentes émotions exprimées. De plus, sur les graphiques 13 à 16, les barres qui correspondent aux émotions exprimées pour lesquelles des jugements d'intensité émotionnelle plus élevés sont prédits sont signalées par un cercle rouge (gris) entourant leurs labels.

Tableau 5: Moyennes et écarts-types par émotion exprimée pour l'intensité perçue de joie, de peur, de tristesse et de colère (N = 18).

intensité émotionnelle perçue	joie		trist		anx		irrit		exalt		desp		paniq		rage	
	moy	sd	moy	sd	moy	sd	moy	sd	moy	sd	moy	sd	moy	sd	moy	sd
joie	4.3	1.6	0.7	0.4	2.0	1.3	2.0	1.2	7.0	1.0	2.1	1.3	3.0	0.9	1.3	0.9
peur	1.1	0.6	2.3	1.4	4.7	1.3	1.2	0.3	2.0	1.3	4.9	1.3	6.5	1.3	1.3	0.5
tristesse	1.8	1.5	6.7	1.1	3.5	1.3	1.3	0.5	2.0	1.0	6.3	1.3	3.4	1.6	1.0	0.4
colère	1.0	0.7	0.9	0.4	2.0	0.8	5.7	1.4	2.2	1.6	3.2	2.0	3.7	1.5	8.3	0.8

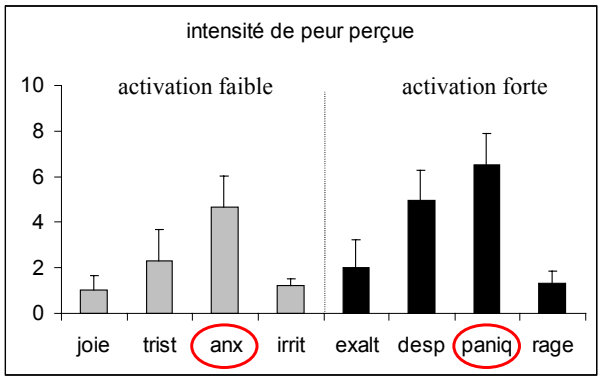
Les jugements moyens d'intensité de joie sont plus élevés pour les expressions de joie intense et de joie calme que pour les autres expressions. De plus, les jugements d'intensité de joie sont plus élevés pour les expressions de joie intense que pour les expressions de joie calme. On observe le même pattern pour les jugements moyens d'intensité de colère; avec des jugements plus élevés pour les expressions de colère chaude et de colère froide que pour les autres expressions et des jugements plus élevés pour les expressions de colère chaude que pour les expressions de colère froide.

Graphique 13: Moyennes et écarts-types pour l'intensité de joie perçue en fonction de l'émotion exprimée



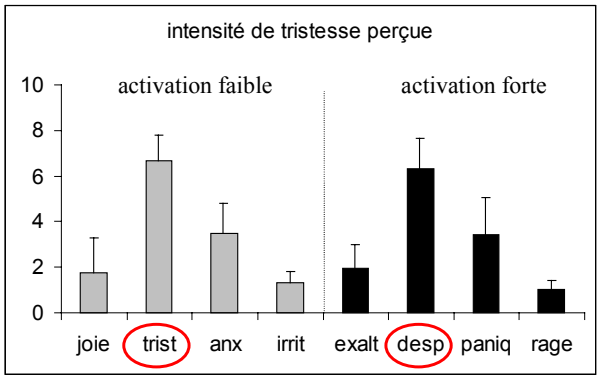
- trist < irrit, anx, desp, paniq, joie, exalt
- rage < paniq, joie, exalt
- irrit, anx, desp, paniq < joie, exalt
- joie < exalt

Graphique 14: Moyennes et écarts-types pour l'intensité de peur perçue en fonction de l'émotion exprimée



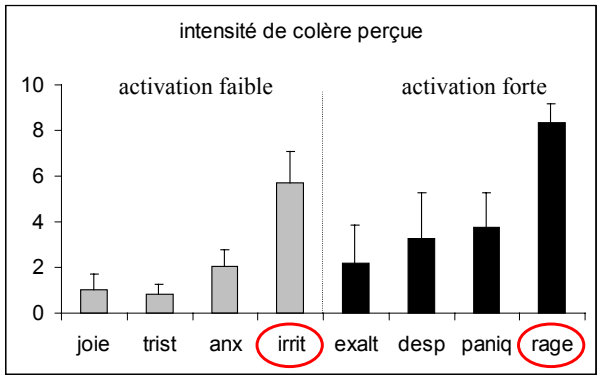
- joie < trist, anx, desp, paniq
- irrit, rage, exalt, trist < anx, desp, paniq
- anx, desp < paniq

Graphique 15: Moyennes et écarts-types pour l'intensité de tristesse perçue en fonction de l'émotion exprimée



- rage, irrit, joie, exalt < paniq, anx, desp, trist
- paniq, anx < desp, trist

Graphique 16: Moyennes et écarts-types pour l'intensité de colère perçue en fonction de l'émotion exprimée



- trist < exalt, desp, paniq, irrit, rage
- joie < desp, paniq, irrit, rage
- anx, exalt < paniq, irrit, rage
- desp, paniq < irrit, rage
- irrit < rage

Le pattern des jugements d'intensité émotionnelle est différent pour l'intensité perçue de peur et tristesse. On observe une "confusion" en ce qui concerne les jugements d'intensité de peur perçue; les expressions de désespoir reçoivent des attributions d'intensité de peur aussi élevées que les attributions effectuées pour les expressions d'anxiété. En revanche, les expressions de peur panique reçoivent des jugements d'intensité de peur plus élevés que les expressions d'anxiété et de désespoir.

Les jugements d'intensité de tristesse sont plus élevés pour les expressions de tristesse et désespoir que pour les autres expressions. L'intensité de tristesse perçue est en revanche égale pour les expressions de tristesse et pour les expressions de désespoir.

Sur les graphiques 13 à 16, les barres grises représentent les jugements d'intensité émotionnelle perçue pour les expressions qui comprennent un faible niveau d'activation émotionnelle, alors que les barres noires représentent les jugements pour les expressions qui comprennent un niveau d'activation émotionnelle élevé. L'examen de ces graphiques révèle que les jugements d'intensité ne sont pas indépendants du niveau d'activation émotionnelle lié aux émotions exprimées. Des ANOVAs à mesures répétées avec un facteur inter/between 'locuteur' (9 niveaux) et trois facteurs intra/within 'activation' (2 niveaux), 'phrase' (2 niveaux) et 'type d'émotion exprimée' (4 niveaux: tristesse/désespoir, colère froide/chaude, joie calme/intense et anxiété/peur panique) confirment la présence d'un effet principal significatif du niveau d'activation émotionnelle sur les jugements d'intensité de peur, de joie et de colère. Les effets principaux du type d'émotion exprimée et du niveau d'activation, ainsi que l'effet d'interaction entre ces deux facteurs sont représentés dans le tableau 6. Les effets de la phrase et des autres interactions sont non-significatifs à l'exception d'un effet principal de la phrase sur l'intensité de tristesse perçue ($F(1, 8) = 10.68$, $p = .011$, $\eta^2 = .57$).

Tableau 6: ANOVAs à mesures répétées – effets de l'activation, du type d'émotion (4 niveaux) et de leur interaction sur l'intensité perçue de joie, peur, tristesse et colère.

émotion perçue	source	df	F	sig.	eta^2
joie	**activation**	**1, 8**	**33.91**	**.000**	**0.81**
	type d'émotion	**3, 24**	**53.70**	**.000**	**0.87**
	activ * type émo	**3, 24**	**18.62**	**.000**	**0.70**
peur	**activation**	**1, 8**	**46.29**	**.000**	**0.85**
	type d'émotion	**3, 24**	**79.40**	**.000**	**0.91**
	activ * type émo	**3, 24**	**6.87**	**.002**	**0.46**
tristesse	activation	1, 8	0.56	.474	0.07
	type d'émotion	**3, 24**	**69.48**	**.000**	**0.90**
	activ * type émo	3, 24	0.38	.769	0.05
colère	**activation**	**1, 8**	**110.25**	**.000**	**0.93**
	type d'émotion	**3, 24**	**134.64**	**.000**	**0.94**
	activ * type émo	3, 24	1.89	.158	0.19

Les expressions qui correspondent à des émotions fortement activées reçoivent des jugements d'intensité de peur, de joie et de colère plus élevés que les expressions qui correspondent à des émotions faiblement activées, indépendamment du locuteur, de la phrase et du type d'émotion exprimée. Les jugements relatifs à l'intensité de tristesse ne sont en revanche pas affectés par le niveau d'activation émotionnelle. Pour les quatre intensités émotionnelles perçues, l'effet du type d'émotion exprimé est cependant aussi très largement indépendant du niveau d'activation sous-

jacent aux émotions exprimées. Contrairement aux résultats obtenus pour les caractéristiques acoustiques (v. section 2.3.2) et les caractéristiques vocales perçues (v. section 3.3.5), l'effet de l'émotion exprimée sur l'intensité émotionnelle perçue ne peut donc pas être attribué dans une large mesure au niveau d'activation émotionnelle.

4.4 Conclusions

L'objectif principal de la procédure utilisée pour récolter les attributions émotionnelles était d'obtenir des jugements continus relatifs à l'intensité émotionnelle perçue. Dans ce contexte, la dimension continue des jugements représente un point particulièrement crucial. En principe, les jugements dichotomiques d'absence/présence pour différents types d'émotions conduisent plus facilement à sélectionner un certain type d'émotion (à l'exclusion des autres types) et relèvent donc d'avantage de la discrimination que de la reconnaissance des émotions exprimées. En pratique, nous avons observé que les auditeurs réalisent également des jugements d'exclusion lorsqu'ils répondent en effectuant des jugements continus. Pour chaque intensité émotionnelle évaluée (intensité de peur, de tristesse, de colère, de joie) une intensité émotionnelle quasi-nulle a été attribuée à une très large majorité des expressions. A notre avis, cet aspect de la distribution des réponses reflète toutefois les caractéristiques réellement perçues des expressions utilisées et ne correspond donc pas à une discrimination réalisée entre différents types d'émotions. Les auditeurs ne sont en effet à aucun moment confrontés à un ensemble d'alternatives de réponses, les échelles émotionnelles leur sont présentées successivement, ce qui ne laisse pas présager des catégories qui pourraient être introduites par la suite. En conséquence, nous postulons que les nombreux jugements d'intensité émotionnelle (quasi-)nuls que nous avons obtenus correspondent bien à l'absence de perception de l'émotion considérée (peur, tristesse, colère ou joie) dans les expressions évaluées.

Le faible nombre d'expressions produit par chaque locuteur pour chaque émotion exprimée (2 expressions pour chacune des 8 catégories émotionnelles) ne permet pas de généraliser les différences observées pour les interactions entre l'émotion exprimée et le locuteur. Il nous semble toutefois intéressant de relever que l'influence du locuteur sur les attributions émotionnelles semble dans une assez forte mesure dépendre du type d'émotion exprimée. De plus, si l'observation que certains locuteurs seraient capables de communiquer avec succès un type d'émotion donné (par exemple la colère) mais échoueraient à communiquer un autre type d'émotion (par exemple la joie) semble théoriquement plausible, il n'existe pas, à notre connaissance, de données publiées à ce sujet. Cette observation mériterait donc d'être répliquée dans un contexte où des attributions émotionnelles seraient disponibles pour un plus grand nombre d'expressions produites par chaque locuteur.

La problématique de la confusion plus systématique entre certains type d'émotions qu'entre d'autres types d'émotions a été décrite en détail pas Banse & Scherer (1996). Ces auteurs ont notamment observé l'existence de plusieurs axes de confusion tels que la valence des expressions (les expressions correspondant à des émotions positives sont plus souvent confondues entre elles qu'avec des expressions correspondant à des émotions négatives) ou le degré d'activation émotionnelle (les expressions correspondant à des émotions comprenant un fort niveau d'activation sont plus souvent confondues entre elles que les expressions correspondant à des émotions comprenant un faible niveau d'activation). Les données que nous avons obtenues ne permettent pas d'effectuer des comparaisons directes avec des études dans lesquelles des proportions de réponses correctes et erronées ont été calculées. En revanche, la problématique de la confusion apparaît également dans nos résultats. Les jugements relatifs à l'intensité de tristesse et les jugements relatifs à l'intensité de peur perçue sont notamment positivement corrélés. L'examen des graphiques 14 et 15 révèle que l'intensité moyenne de peur perçue et aussi importante pour les expressions de désespoir que pour les expressions d'anxiété, en revanche les expressions de tristesse reçoivent des jugements d'intensité de peur plus faibles que les expressions d'anxiété. En d'autres termes, les expressions de désespoir produites par les acteurs communiquent de la tristesse (v. graphique 15) mais également de la peur (v. graphique 16), ce qui n'est pas le cas des expressions de tristesse qui communiquent une intensité de tristesse élevée et une intensité de peur faible. De plus, la "confusion" entre peur et tristesse (pour les expressions de désespoir) semble ne pas être réciproque au sens où les expressions d'anxiété et les expressions de peur panique reçoivent des jugements d'intensité de tristesse significativement plus faibles que les expressions de tristesse et de désespoir. Cependant, on observe également que les expressions d'anxiété et de peur panique communiquent une intensité de tristesse plus importante que les expressions de joie calme et forte et que les expressions de colère chaude et froide; ce qui indique, à notre avis, que les expressions d'anxiété et de peur panique pourraient être plus aisément "confondues" avec des expressions de tristesse que les autres types d'expressions.

D'autre part, les résultats présentés ont permis de mettre en évidence la différence entre l'activation émotionnelle et l'intensité émotionnelle qui avait été préalablement formulée sur le plan théorique. Les jugements relatifs à l'intensité de tristesse perçue – qui s'étendent de l'absence de tristesse (intensité nulle) à la présence d'une tristesse extrêmement forte – ne sont pas liés au niveau d'activation émotionnelle. Les expressions de tristesse (qui comprennent un niveau d'activation faible) ne sont notamment pas perçues comme exprimant une intensité de tristesse plus faible que les expressions de désespoir (qui comprennent un niveau d'activation élevé). En revanche, une association entre l'activation et l'intensité est apparue pour les jugements de colère, de peur et de

joie. Pour les jugements de colère, les expressions fortement actives correspondant à la colère chaude sont perçues comme exprimant une intensité de colère plus forte que les expressions faiblement actives correspondant à la colère froide. La même conclusion s'applique à l'intensité de joie perçue – plus forte pour les expressions de joie intense que pour les expressions de joie calme – et à l'intensité de peur perçue – plus forte pour les expressions de peur panique que pour les expressions d'anxiété. A notre avis, l'absence de relation entre le niveau d'activation de l'émotion exprimée et l'intensité de tristesse perçue illustre le fait que la relation qui existe plus globalement entre le niveau d'activation physiologique associé aux émotions et l'intensité des réactions émotionnelles vécues n'est pas absolue. La tristesse dans sa version "déprimée" représente un exemple frappant d'une d'émotion qui peut être subjectivement perçue comme correspondant une émotion très forte alors même qu'elle comporte une composante d'activation physiologique très faible.

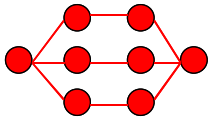
Des explications alternatives à l'absence de relation entre le niveau d'activation et l'intensité émotionnelle perçue pour la tristesse peuvent évidemment être formulées. Il est possible, par exemple, d'argumenter que la tristesse "déprimée" correspond à la version la plus prototypique de la tristesse. En conséquence, la tristesse "désespérée" ne serait pas aussi facilement qualifiée de 'tristesse' et recevrait des attributions d'intensité de 'tristesse' comparativement plus faibles. Pour contrer cet argument, on peut toutefois relever que les expressions de désespoir reçoivent un jugement moyen de 6.3 sur l'échelle d'intensité de tristesse, le désespoir – exprimé vocalement – serait donc bien identifié comme une forme de 'tristesse'. En revanche, à l'appui de l'argument de l'appartenance non-univoque du désespoir à la catégorie 'tristesse', on peut avancer que les expressions de désespoir reçoivent également un jugement moyen de 4.9 – soit un jugement relativement élevé – sur l'échelle d'intensité de peur. Ceci suggère, soit que les expressions de désespoir sont identifiées comme telles et que le désespoir pourrait éventuellement comporter à la fois une composante de tristesse et une composante de peur sur le plan de sa description subjective, soit que les expressions de désespoir ne sont pas identifiées comme telles et sont perçues comme exprimant alternativement une forme de tristesse ou une forme de peur.

Dans un prolongement de cette étude, il serait intéressant d'obtenir des données relativement à la manière dont les expressions de désespoir que nous avons utilisées seraient catégorisées au cas où des alternatives de réponse 'peur', 'désespoir' et 'tristesse' (au sens de 'tristesse déprimée') seraient proposées. Dans l'hypothèse où les expressions de désespoir seraient discriminées correctement (c'est-à-dire distinguées des catégories 'tristesse déprimée' et 'peur'), il serait possible de postuler que la "confusion" que nous avons observée pour les expressions de désespoir (qui reçoivent des attributions d'intensité de tristesse et de peur relativement élevées) serait essentiellement due au fait

que le 'désespoir' peut être décrit (notamment lorsque cette catégorie n'est pas disponible) comme un mélange de peur et de tristesse.

Certains aspects de la différence qui peut exister entre l'activation sous-jacente aux émotions exprimées et l'intensité émotionnelle perçue ont pu être examiné dans cette étude. En revanche, nous n'avons pas d'indication relativement à l'activation émotionnelle perçue par les auditeurs. Nous avons toutefois constaté dans les chapitres précédents que le niveau d'activation sous-jacent aux émotions exprimées est associé à un ensemble de caractéristiques vocales. Les expressions fortement actives sont notamment caractérisées, aussi bien sur le plan acoustique que sur le plan des caractéristiques vocales perçues, par une augmentation de l'énergie, de la rapidité et de la hauteur relativement aux expressions faiblement actives. A notre avis, le niveau d'activation serait donc facilement identifiable pour les expressions émotionnelles que nous avons utilisées et se distinguerait du niveau d'intensité perçu. L'activation perçue des expressions de désespoir serait notamment plus importante que l'activation perçue pour les expressions de tristesse, contrairement à l'intensité émotionnelle perçue qui est aussi importante pour les expressions de tristesse que pour les expressions de désespoir. Dans un autre prolongement suggéré de cette étude, cette hypothèse pourrait être confirmée ou invalidée en demandant à un nouveau groupe d'auditeurs de juger l'activation émotionnelle sous-jacente aux expressions vocale en utilisant la même procédure que pour l'évaluation de l'intensité émotionnelle perçue.

5 Modéliser le processus de communication vocale



5.1 Introduction

Après avoir examiné successivement dans les chapitres précédents les caractéristiques acoustiques, les caractéristiques vocales perçues et les émotions attribuées pour un ensemble d'expressions vocales émotionnelles, nous allons présenter, dans la section qui suit, une modélisation de l'ensemble du processus de la communication vocale des émotions. Cette modélisation a pour objectif d'inclure simultanément le versant de l'encodage et le versant du décodage des expressions vocales émotionnelles, en partant des expressions émotionnelles encodées par des locuteurs-émetteurs, en passant par les caractéristiques acoustiques et les caractéristiques vocales perçues des expressions pour aboutir à l'attribution émotionnelle réalisée par des auditeurs-récepteurs.

Cette dernière étape de la présentation des résultats se base sur la proposition de Scherer (1978, 2003) d'utiliser une version modifiée du "modèle en lentille" (lens model) de Brunswik comme paradigme pour l'étude de la communication non-verbale. Le paradigme suggéré par Scherer, ainsi que l'utilisation proposée du "modèle en lentille" par d'autres auteurs dans le domaine de la communication non-verbale ont été décrits dans l'introduction générale de cette thèse (section 1.3). Dans la section qui suit, deux modèles statistiques ont été retenus pour l'opérationnalisation du "modèle en lentille" de Brunswik: d'une part, le modèle utilisé par Juslin (1998) pour ses études de la communication des émotions par la musique et, d'autre part, le modèle utilisé par Scherer (1978) dans son étude de la communication des traits de personnalité par la voix.

Le modèle statistique employé par Juslin (1998) propose une décomposition de la corrélation entre l'émotion exprimée et l'émotion perçue – qui correspond au degré de *réussite* de la communication, *achievement* ou *accuracy* dans la terminologie brunswikienne – sous forme d'une équation ("Lens Model Equation"). Cette équation divise le degré de *réussite* (*achievement*) de la communication en quatre composantes: (1) la *validité écologique* du modèle (*ecological validity*) qui correspond à la corrélation multiple entre l'émotion exprimée et les caractéristiques vocales (acoustiques ou perçues) introduites dans le modèle, (2) la *validité fonctionnelle* du modèle (*functional validity*) qui correspond à la corrélation multiple entre l'émotion perçue et les caractéristiques vocales introduites dans le modèle, (3) la correspondance (*matching*) entre les deux parties du modèle, c'est-à-dire la corrélation entre la prédiction de l'émotion exprimée et la prédiction de l'émotion perçue par les caractéristiques vocales introduites dans le modèle et, finalement, (4) la partie de la *réussite* (*achievement*) qui n'est pas expliquée par les caractéristiques vocales introduites dans le modèle.

La deuxième approche statistique, reprise de Scherer (1978), exploite un modèle simple d'équations structurales. Dans cette approche, le modèle testé distingue les caractéristiques vocales *distales* (qui correspondent aux paramètres acoustiques dans notre étude) des caractéristiques vocales *proximales* (qui correspondent aux caractéristiques vocales perçues dans notre étude). Une équation structurale (path analysis) permet de décomposer la corrélation entre l'émotion exprimée et l'émotion perçue (*achievement/accuracy*) entre les différents "paths" qui relient l'émotion exprimée à l'émotion perçue dans le modèle. Le modèle inclut 4 types de "paths": (1) les *effets centraux* (*central effects*) qui correspondent à la part de la communication modélisée par les caractéristiques vocales distales et proximales, (2) les *effets basés sur les indices distaux* (*distally based effects*), c'est-à-dire la part de la communication modélisée seulement par les caractéristiques vocales distales (les paramètres acoustiques), (3) les *effets basés sur les indices proximaux* (*proximally based effects*), qui correspondent à la part de la communication modélisée seulement par les caractéristiques vocales proximales (les caractéristiques vocales perçues), (4) l'*effet direct* (*direct effect*) entre l'émotion exprimée et l'émotion perçue, c'est-à-dire la part de la communication qui n'est pas expliquée par les caractéristiques vocales distales ou proximales qui figurent dans le modèle.

Nous avons choisi d'utiliser ces deux modèles dans la mesure où ils ont été précédemment appliqués à des questions voisines de la question examinée dans cette thèse. Les deux modèles sont fondés sur des analyses corrélationnelles (des régressions multiples), mais ils proposent un traitement des données qui accorde plus ou moins d'importance à différents paramètres des modèles de régression et, comme nous venons de l'exposer ci-dessus, ils divisent le processus de communication en composantes différentes. L'utilisation simultanée de deux modèles statistiques présente en outre l'avantage de relativiser chacun des modèles en mettant en évidence d'éventuelles différences entre les résultats obtenus par chacun des modèles pour les mêmes données.

Les deux modèles statistiques employés seront présentés plus en détails ci-dessous. Auparavant, deux problématiques qui seront adressées dans les analyses présentées dans cette section méritent, selon nous, d'être mises en évidence. La première problématique a trait à l'objectif même de cette section de présentation des données qui vise à représenter dans une même analyse les deux pôles (encodage/production et décodage/perception) du processus de communication. La seconde problématique abordée ci-dessous a trait au rôle de l'activation dans les descriptions – ou dans les modèles – habituellement proposé(e)s de la communication vocale des émotions.

5.1.1 Considérer simultanément l'encodage et le décodage des expressions vocales

La production (l'encodage) et la perception (le décodage) des expressions vocales émotionnelles ont été le plus souvent considérées séparément dans des études qui se sont focalisées soit sur la

description des caractéristiques acoustiques correspondant à des émotions exprimées, soit sur la capacité à reconnaître différents types d'émotions. En fait, il serait plus correct de dire que ces deux versants de la communication ont été – en apparence seulement – étudiés séparément. En réalité, ils interviennent simultanément dans la plupart des études réalisées qui ne distinguent pas assez clairement l'influence réciproque des deux pôles de la communication sur leurs résultats.

Ainsi, les études de production incluent souvent l'aspect de la reconnaissance en sélectionnant des expressions vocales qui sont clairement discriminées par un groupe d'auditeurs comme correspondant à différents types d'émotions. Cette pratique vise explicitement à ne pas se fier uniquement à la compétence des encodeurs (qui pourraient produire des expressions qui ne correspondraient pas nécessairement aux émotions ciblées); elle a pour objectif de contrôler la "qualité" des expressions vocales produites afin de n'inclure que des expressions susceptibles de communiquer les émotions ciblées. Cette pratique en principe louable a pour inconvénient d'introduire au moins deux sources de confusion. Premièrement, les résultats qui sont obtenus dans ces études (les profils acoustiques qui seront définis) correspondent d'avantage à des caractéristiques perçues comme émotionnelles qu'à des émotions exprimées. Ce point n'est que rarement explicite; la sélection par la reconnaissance est souvent "mise de côté" et les résultats sont présentés comme des profils acoustiques correspondant aux différents types d'émotions exprimées. Deuxièmement, la compétence des décodeurs étant variable au même titre que la compétence des encodeurs, il se pourrait qu'un groupe d'auditeurs ne discrimine pas des expressions qui seraient malgré cela représentatives d'un certain type d'émotion vécu et/ou exprimé par un encodeur. En d'autres termes, il serait incorrect de considérer a priori que le critère de l'émotion perçue/reconnue est plus conforme à la "réalité" de l'émotion encodée que le critère de l'intention expressive de l'encodeur (ou de l'émotion vécue par l'encodeur).

Les études de la reconnaissance (décodage) quant à elles incluent également le versant de l'encodage, au sens où elles déterminent des taux de reconnaissance correcte pour différents types *d'émotions exprimées*. Aussi, dans ces études, lorsque les émotions théoriquement exprimées dans des enregistrements vocaux sont mal discriminées, il n'est pas possible d'identifier la source de l'échec de la communication. Il pourrait s'agir d'une "faillite au décodage" – l'émotion ciblée est bien exprimée mais n'est pas reconnue – comme d'une "faillite à l'encodage" – l'émotion ciblée n'a pas été encodée par le(s) locuteur(s).

Pour ces différentes raisons, il nous semble important de distinguer les deux versants (encodage-décodage) de la communication vocale des émotions et de les inclure conjointement et de manière

explicite dans les études réalisées dans ce domaine; ceci constitue l'objectif principal de l'analyse des résultats présentée dans cette quatrième et dernière étape de présentation des résultats.

5.1.2 Contrôler l'influence du niveau d'activation

La seconde problématique que nous avons choisi de développer avant de revenir aux modèles statistiques concerne l'influence du niveau d'activation émotionnelle sur les résultats qui sont habituellement obtenus dans les études qui traitent de la communication vocale émotionnelle. Les revues de la littérature dans ce domaine soulignent régulièrement que les caractéristiques vocales – en général évaluées sur le plan acoustique – semblent refléter avant tout le niveau d'activation émotionnelle sous-jacent aux émotions exprimées. Nous avons nous même observé dans les sections précédentes que de nombreuses caractéristiques que nous avons mesurées (sur plan acoustiques et sur le plan des caractéristiques vocales perçues) semblent effectivement varier de manière importante en fonction du niveau d'activation sous-jacent aux états émotionnels exprimés.

En conséquence, il apparaît que la comparaison entre des expressions qui correspondent à différents types d'émotions – par exemple la colère et la tristesse – ne devrait pas recouvrir simultanément différents niveaux d'activation émotionnelle, sous peine de ne pas pouvoir départager l'effet du type d'émotion exprimé de l'effet de l'activation sous-jacente. La confusion entre le type d'émotion et le niveau d'activation pourrait par exemple conduire à conclure que la tristesse présente une intensité vocale plus faible que la colère alors que l'intensité vocale pourrait être liée directement au niveau d'activation sous-jacent qui est plus faible pour la tristesse que pour la colère.

La comparaison entre des expressions émotionnelles et des expressions "neutres" qui est parfois effectuée pose le même problème. Les expressions "neutres" ne sont réellement neutres que sur le plan de la valence du sentiment subjectif (l'adjectif "neutre" qualifiant en général un état qui n'est vécu comme ni positif, ni négatif). Sur le plan de l'activation, l'état neutre est toutefois plus proche d'une émotion telle que la tristesse (peu activée) que d'une émotion telle que la colère (très activée). Ainsi la différence entre une expression "neutre" et une expression de colère pourrait refléter une différence liée à l'activation sous-jacente à ces deux états; alors que la différence entre une expression neutre et une expression de tristesse ne sera pas aussi fortement affectée par le niveau d'activation qui est moins variable pour ces deux états.

La présentation des résultats proposée ci-dessous contrôle l'influence du niveau d'activation sur les résultats tout en comparant quatre types d'émotions: la joie, la peur, la colère et la tristesse. Pour chaque type d'émotion, la moitié des expressions utilisées comportent un niveau d'activation faible et l'autre moitié un niveau d'activation fort. Les expressions correspondant à chacun des types d'émotions – comprenant 50% d'expressions faiblement activées et 50% d'expression fortement

activées – sont comparées aux trois autres types d'émotions qui comprennent également 50% d'expressions faiblement activées et 50% d'expression fortement activées. Pour l'interprétation des résultats présentés, il importe donc de garder à l'esprit que, contrairement aux résultats habituellement proposés dans la littérature qui comparent souvent des expressions de tristesse (faiblement activées) avec des expressions de colère chaude ou de peur panique (fortement activées), les résultats présentés ci-dessous pour chaque type d'émotion ne reflètent en principe pas un effet attribuable à l'activation sous-jacente aux émotions considérées. L'effet de l'activation est par ailleurs directement examiné en opposant les expressions de peur, de tristesse, de colère et de joie faiblement activées aux expressions de peur, de tristesse, de colère et de joie fortement activées.

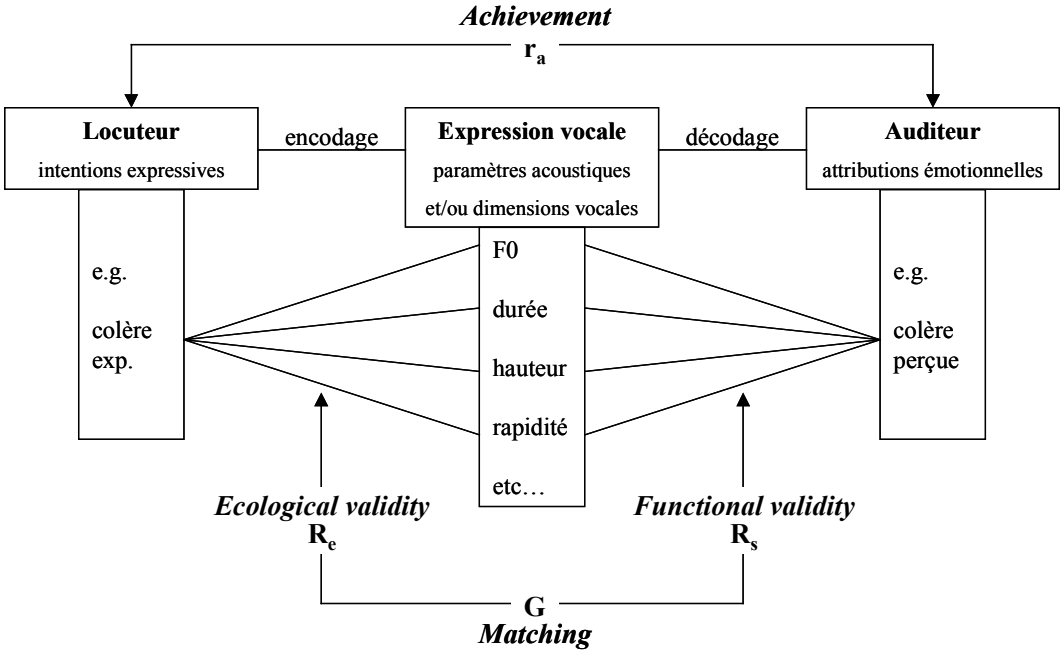
5.2 Modèles

Dans cette section, les deux modèles statistiques retenus pour l'analyse des résultats seront décrits en détails. Une partie de la terminologie utilisée par les deux auteurs qui sont à l'origine de ces modèles figure en Anglais dans le texte et les illustrations. L'absence de traduction pour ces termes constitue un choix délibéré qui découle de la difficulté à trouver des termes français qui recouvrent exactement la signification des termes originaux anglais. Le modèle repris de Juslin (1998) sera présenté dans un premier temps. Le modèle proposé par Scherer (1978) sera présenté ensuite.

5.2.1 Lens Model Equation (proposition de Juslin, 1998)

Dans sa thèse de 1998, Juslin a présenté une adaptation du modèle en lentille de Brunswik pour l'étude de la communication des émotions par la musique. La figure 1 correspond à une nouvelle adaptation de ce modèle au contexte de l'étude de la communication vocale des émotions. Cette figure représente quatre aspects de la communication. Premièrement, la relation entre l'émotion exprimée et l'émotions perçue (*achievement*, r_a) qui correspond au degré de "réussite" de la communication. Deuxièmement, la relation entre un ensemble de caractéristiques vocales (acoustiques et/ou perçues) et l'émotion exprimée (*ecological validity*, R_e) qui indique à quel point les paramètres acoustiques et/ou les aspects vocaux perçus qui se trouvent dans le modèle caractérisent l'émotion exprimée. Troisièmement, la relation entre les caractéristiques vocales et l'émotion perçue (*functional validity*, R_s) qui indique à quel degré l'émotion perçue est caractérisée par les paramètres acoustiques et/ou les aspects vocaux perçus qui se trouvent dans le modèle. Et quatrièmement, la relation entre les deux côtés du modèle (*matching*, G) qui indique à quel point la relation observée entre les caractéristiques vocales et l'émotion exprimée est similaire à la relation observée entre les caractéristiques vocales et l'émotion perçue. En d'autres termes, le *matching* indique dans quelle mesure l'utilisation des caractéristiques vocales à l'encodage correspond à l'utilisation des caractéristiques vocales au décodage.

Figure 1: Modèle en lentille proposé par Juslin (1998), adapté pour la communication vocale des émotions



Sur le plan statistique, Juslin a appliqué à l'étude de la communication des émotions par la musique un modèle qui a été originalement développé par (Hursh, Hammond, & Hursh, 1964) et (Tucker, 1964). Ce modèle correspond à une équation (Lens Model Equation ou LME, v. équation 1 ci-dessous) qui démontre que l'*achievement* de la communication (r_a, la corrélation entre les émotions exprimées et les émotions perçue) correspond à la somme de deux composantes. Une composante désignée comme la *composante linéaire* (*linear component*) et une *composante non-modélisée* (*unmodeled component*).

$$r_a = G\, R_e R_s + C\sqrt{1 - R_e^2}\sqrt{1 - R_s^2} \qquad (1)$$

La *composante linéaire* est le produit de la *cohérence de l'encodage* (R_e, *performer consistency* dans la terminologie de Juslin qui étudie l'encodage de l'émotion effectué par des musiciens), de la *cohérence du décodage* (R_s, *listener consistency*) et de la *correspondance* (G, *matching*) entre l'encodage et le décodage. La valeur R_e (*cohérence de l'encodage*) correspond à la corrélation multiple entre les caractéristiques vocales qui figurent dans le modèle et l'émotion exprimée. La valeur R_s (*cohérence du décodage*) correspond à la corrélation multiple entre les caractéristiques vocales qui figurent dans le modèle et l'émotion perçue. La valeur G (*matching*) correspond à la corrélation entre la prédiction de l'émotion exprimée et la prédiction de l'émotion perçue; les deux variables prédites sont dérivées de la régression multiple des caractéristiques vocales sur l'émotion exprimée, d'une part, et sur l'émotion perçue, d'autre part.

La composante *non-modélisée* est également le produit de 3 paramètres: (1) la racine carrée de la variance résiduelle de l'émotion exprimée – après soustraction de la variance prédite par les caractéristiques acoustiques qui figurent dans le modèle, soit $\sqrt{1-R_e^2}$ – (2) la racine carrée de la variance résiduelle de l'émotion perçue – après soustraction de la variance prédite par les caractéristiques acoustiques qui figurent dans le modèle, soit $\sqrt{1-R_s^2}$ – (3) la corrélation entre les résidus de l'émotion exprimée et les résidus de l'émotion perçue (C); les deux variables résiduelles sont dérivées de la régression multiple des caractéristiques vocales sur l'émotion exprimée, d'une part, et sur l'émotion perçue, d'autre part.

Pour l'interprétation des résultats, on peut considérer que la plupart des paramètres dans cette équation se comportent comme des corrélations. L'élévation au carré de la corrélation multiple R_e correspond ainsi à la part de variance de l'émotion exprimée qui peut être expliquée par les caractéristiques acoustiques qui figurent dans le modèle. De même, l'élévation au carré de la corrélation multiple R_s correspond à la part de variance de l'émotion perçue qui peut être expliquée par les caractéristiques acoustiques qui figurent dans le modèle. L'élévation au carré de la corrélation r_a correspond à la part de variance partagée entre l'émotion exprimée et l'émotion perçue, alors que l'élévation au carré de la corrélation G correspond à la part de variance partagée entre l'émotion exprimée et l'émotion perçue prédites par les caractéristiques acoustiques qui figurent dans le modèle. Il importe aussi de relever que si, par définition, les paramètres $1-R_e^2$ et $1-R_s^2$ sont complémentaires respectivement des paramètres R_e^2 et R_s^2, le paramètre G^2 n'est pas complémentaire du paramètre C^2.

Dans ce modèle, une valeur proche de 1 pour le paramètre G indique une bonne *correspondance* (*matching*) au niveau de l'utilisation des caractéristiques vocales des deux côtés du modèle, une valeur proche de 0 pour ce paramètre indiquerait au contraire que l'utilisation des caractéristiques vocales est différente à l'encodage et au décodage. Des valeurs élevées pour les paramètres $\sqrt{1-R_e^2}$ et $\sqrt{1-R_s^2}$ – c'est-à-dire par définition des valeurs faibles pour les paramètres R_e^2 et R_s^2 – peuvent être la conséquence de plusieurs facteurs que le modèle ne permet pas de départager: (1) Les caractéristiques vocales qui figurent dans le modèle sont utilisées de manière "inconstante" (incohérente). (2) Les caractéristiques vocales qui figurent dans le modèle sont utilisées de manière non linéaire – certaines configurations ou des fonctions non-linéaires de ces caractéristiques permettraient de prédire l'émotion exprimée et l'émotion perçue, alors que des combinaisons linéaires ne le permettent pas. (3) Des caractéristiques vocales importante pour l'encodage ou le

décodage ne sont pas inclues dans le modèle. (4) Les erreurs de mesure sont importantes pour les variables considérées – l'émotion exprimée ou perçue et/ou les caractéristiques vocales ne sont pas mesurées avec suffisamment de précision.

De même une valeur élevée (proche de 1) pour le paramètre C peut signaler différents problèmes que le modèle ne permet pas de départager: (1) Des caractéristiques vocales qui sont utilisées de manière similaire à l'encodage et au décodage ne sont pas inclues dans le modèle. (2) Les caractéristiques vocales qui figurent dans le modèle sont utilisées en configurations similaires – ou plus généralement de la même manière non-linéaire – à l'encodage et au décodage. (3) La corrélation entre les résidus est due au hasard (cette possibilité est à considérer particulièrement lorsque l'échantillon est faible).

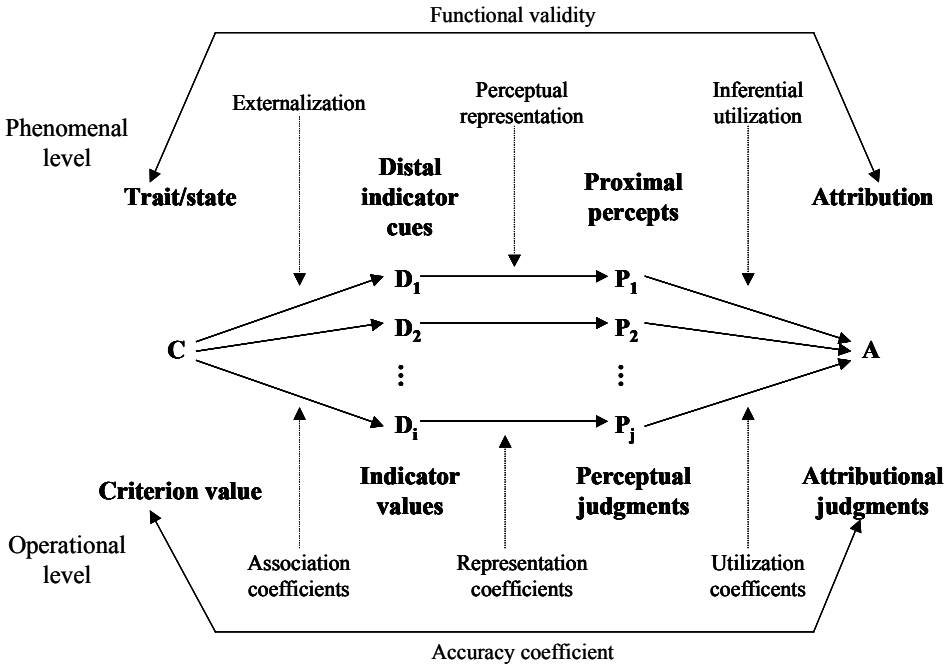
5.2.2 Path analysis (proposition de Scherer, 1978)

Scherer (1978) a appliqué une autre adaptation du modèle en lentille de Brunswik à l'étude de la communication des traits de personnalité par la voix. La figure 2 représente le modèle en lentille proposé par Scherer avec la terminologie employée par cet auteur. Ce modèle décompose le processus de communication en quatre étapes: (1) l'*état interne* (*trait/state*) qui correspond dans notre étude à l'émotion exprimée; (2) les *indices distaux* (*distal indicator cues*) qui correspondent dans notre étude aux caractéristiques vocales acoustiques; (3) les *indices proximaux* (*proximal percepts*) qui correspondent ici aux caractéristiques vocales perçues; et (4) l'*attribution* relative à l'état interne qui correspond dans notre étude à l'émotion perçue. Sur la figure 2, quatre types de relations sont par ailleurs représentées: (1) La relation directe entre l'émotion exprimée et l'émotion perçue – dans la terminologie proposée par Scherer, cette relation correspond à la *validité fonctionnelle* (*functional validity*[16]) du modèle sur le plan du phénomène et à un *coefficient d'exactitude* (*accuracy coefficient*) sur le plan opérationnel. (2) La relation entre l'émotion exprimée et les caractéristiques vocales acoustiques – soit l'*externalisation* qui est représentée par des *coefficients d'association* sur le plan opérationnel. (3) La relation entre les caractéristiques vocales acoustiques et les caractéristiques vocales perçues – cette relation correspond à des *coefficients de représentation* sur le plan opérationnel. (4) La relation entre l'émotion perçue et les caractéristiques vocales perçues – soit l'*inférence* de l'émotion qui correspond, sur le plan opérationnel, à des *coefficients d'utilisation*.

---

[16] Juslin utilise le terme *functional validity* pour qualifier la relation entre les caractéristiques vocales et l'émotion perçue. Pour lui, la "fonctionnalité" de la communication se situerait donc dans une relation forte entre des caractéristiques vocales et l'émotion perçue (indépendamment de la relation entre les caractéristiques vocales et l'émotion perçue, et indépendamment de la correspondance entre l'encodage et le décodage). En revanche, pour Scherer, la "fonctionnalité" de la communication se situe au niveau de l'*achievement/accuracy* (réussite) de la communication.

Figure 2: Modèle en lentille proposé par Scherer (1978)



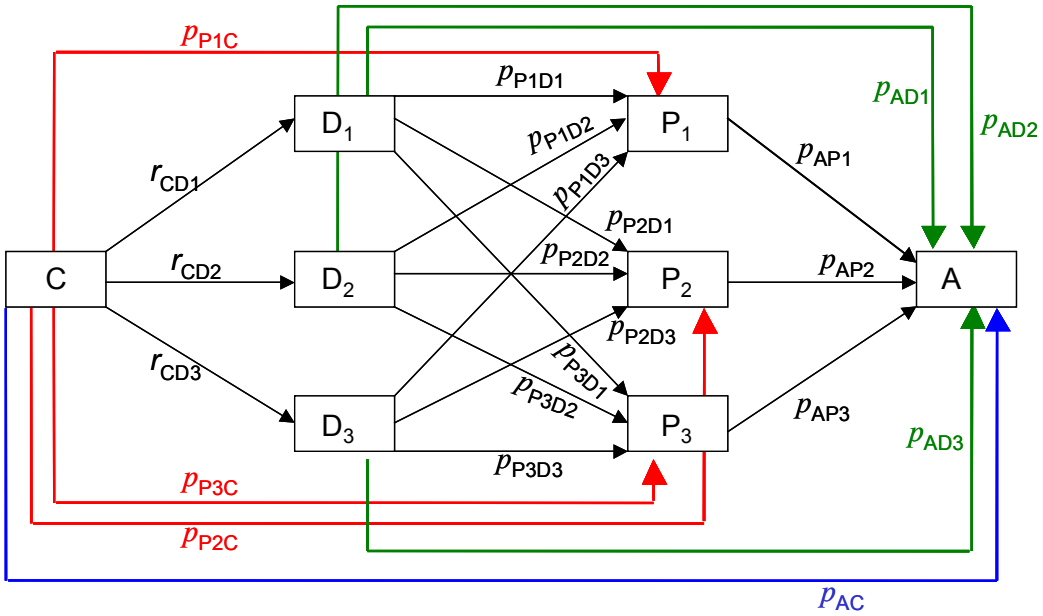
Sur le plan statistique, Scherer utilise un modèle d'équation structurale ("path analysis") représenté par l'équation 2 qui met en jeu un plus grand nombre de paramètres. Cette équation démontre que la corrélation entre l'émotion exprimée et l'émotion perçue (*accuracy coefficient*, r_{CA}) correspond à la somme de 4 composantes. Premièrement, les *effets centraux* (*central effects*, $r_{CD}\,\rho_{PD}\,\rho_{AP}$) qui correspondent à la part de la relation entre l'émotion exprimée et l'émotion perçue qui est médiatisée dans le modèle par les caractéristiques acoustiques (distales) et par les caractéristiques vocales perçues (proximales). Deuxièmement, les *effets périphériques basés sur les indices distaux* (*distally based peripheral effects*, $r_{CD}\,\rho_{AD}$) qui correspondent à la part de la relation entre l'émotion exprimée et l'émotion perçue qui n'est médiatisée que par les caractéristiques acoustiques. Troisièmement, les *effets périphériques basés sur les indices proximaux* (*proximally based peripheral effects*, $\rho_{PC}\,\rho_{AP}$) qui correspondent à la part de la relation entre l'émotion exprimée et l'émotion perçue qui n'est médiatisée que par les caractéristiques vocales perçues. Quatrièmement, l'*effet direct* (*direct effect*, ρ_{AC}) qui correspond à la relation entre l'émotion exprimée et l'émotion perçue qui n'est pas médiatisée par les caractéristiques vocales acoustiques et perçues qui figurent dans le modèle.

$$r_{CA} = r_{CD}\,\rho_{PD}\,\rho_{AP} + r_{CD}\,\rho_{AD} + \rho_{PC}\,\rho_{AP} + \rho_{AC} \qquad (2)$$

Seuls une partie des *effets centraux* et l'*effet direct* sont représentés par des flèches continues sur la figure 2; la figure 3 fournit une illustration plus conforme au modèle statistique proposé par Scherer. Cette illustration représente le cas où trois indices distaux (caractéristiques acoustiques) et

trois indices proximaux (caractéristiques vocales perçues) figurent dans le modèle. Les *effets centraux* sont représentés par une succession de trois flèches noires de l'émotion exprimée à l'émotion perçue. Les *effets périphériques basés sur les indices distaux* sont représentés par une flèche noire, suivie d'une flèche verte. Les *effets périphériques basés sur les indices proximaux* sont représentés par une flèche rouge suivie d'une flèche noire. L'*effet direct* est représenté par la flèche bleue. Le tableau 1 présente le développement de l'équation 2 pour cette situation.

Figure 3: Illustration du modèle statistique ("path analysis") proposé par Scherer, trois indices distaux et trois indices proximaux figurent dans ce modèle



Les paramètres r_{CA} et r_{CD} correspondent à des corrélations entre l'émotion exprimée et l'émotion perçue (r_{CA}) et entre l'émotion exprimée et chacune des caractéristiques acoustiques (r_{CD}). Les autres paramètres correspondent à des coefficients de régression standardisés (betas) dérivés d'une série de régressions multiples (linéaires, non-hiérarchiques). Les coefficients ρ_{AP}, ρ_{AD} et ρ_{AC} sont obtenus en régressant les caractéristiques vocales perçues (ρ_{AP}) et les caractéristiques acoustiques (ρ_{AD}) qui figurent dans le modèle, ainsi que l'émotion exprimée (ρ_{AC}) sur l'émotion perçue. Les coefficients ρ_{PD} et ρ_{PC} sont obtenus en régressant les caractéristiques acoustiques (ρ_{PD}) qui figurent dans le modèle ainsi que l'émotion exprimée (ρ_{PC}) sur chacune des caractéristiques vocales perçues qui figurent dans le modèle.

Tableau 1: Développement de l'équation proposée par Scherer pour le cas illustré par la figure 3

Accuracy coefficient	Central effect	Peripheral effects: Distally based	Peripheral effects: Proximally based	Peripheral effects: Direct effect
r_{CA} =	$\sum_{i,j=1}^{3} r_{CD_i}\rho_{P_jD_i}\rho_{AP_j}$	+ $\sum_{i=1}^{3} r_{CD_i}\rho_{AD_i}$	+ $\sum_{j=1}^{3} \rho_{P_jC}\rho_{AP_j}$	+ ρ_{AC}

r_{CA} =	$r_{CD_1}\rho_{P_1D_1}\rho_{AP_1} + r_{CD_1}\rho_{P_2D_1}\rho_{AP_2} + r_{CD_1}\rho_{P_3D_1}\rho_{AP_3}$	
	$+ r_{CD_2}\rho_{P_1D_2}\rho_{AP_1} + r_{CD_2}\rho_{P_2D_2}\rho_{AP_2} + r_{CD_2}\rho_{P_3D_2}\rho_{AP_3}$	*central effect*
	$+ r_{CD_3}\rho_{P_1D_3}\rho_{AP_1} + r_{CD_3}\rho_{P_2D_3}\rho_{AP_2} + r_{CD_3}\rho_{P_3D_3}\rho_{AP_3}$	
	$+ r_{CD_1}\rho_{AD_1} + r_{CD_2}\rho_{AD_2} + r_{CD_3}\rho_{AD_3}$	*distally based*
	$+ \rho_{P_1C}\rho_{AP_1} + \rho_{P_2C}\rho_{AP_2} + \rho_{P_3C}\rho_{AP_3}$	*proximally based*
	$+ \rho_{AC}$	*direct effect*

Dans le cadre de ce modèle, on souhaite que les caractéristiques vocales (acoustiques et perçues) puissent rendre compte de la relation entre l'émotion exprimée et l'émotion perçue. L'*effet direct* doit donc être aussi faible que possible (le coefficient ρ_{AC} devrait être proche de 0). Un *effet direct* élevé (un coefficient ρ_{AC} proche de la valeur de la corrélation r_{CA}) indiquerait que les caractéristiques vocales qui figurent dans le modèle ne parviennent pas à rendre compte de la relation entre l'émotion exprimée et l'émotion perçue. Cette situation peut être la conséquence de plusieurs facteurs que le modèle ne permet pas de départager: (1) Des caractéristiques vocales qui interviennent dans la communication vocale des émotions ne figurent pas dans le modèle. (2) Les caractéristiques qui figurent dans le modèle interviennent dans la communication de manière non-linéaire ou en configurations. (3) Des erreurs de mesure trop importantes interviennent sur une partie ou sur la totalité des variables dans le modèle.

Plus généralement, un modèle idéal correspondrait au cas où la corrélation entre l'émotion exprimée et l'émotion perçue serait expliquée entièrement par les *effets centraux* ($r_{CA} = r_{CD}\,\rho_{PD}\,\rho_{AP}$). En effet, des *effets périphériques* importants *basés sur les indices proximaux* ou *basés sur les indices distaux* indiquent également que des indices distaux ou des indices proximaux qui jouent un rôle important dans la communication vocale des émotions ne sont pas modélisés. Ceci pourrait être dû au fait que des caractéristiques vocales importantes – distales/acoustiques dans le cas où le paramètre $\rho_{PC}\,\rho_{AP}$ est élevé, ou proximales/perçues dans le cas où le paramètre $r_{CD}\,\rho_{AD}$ est élevé – ne figurent pas dans le modèle ou sont mal représentées (du fait que leur relation avec l'émotion perçue et/ou exprimée n'est pas linéaire ou qu'elle dépende de la configuration avec d'autres caractéristiques, ou encore à cause d'erreurs trop importantes dans la mesure de ces variables).

5.3 Application aux données

Les données qui seront utilisées dans cette section ont été présentées en détails dans les chapitres précédents. En conséquence, seul un bref rappel concernant les variables examinées est présenté ci-dessous.

Des expressions correspondant à 8 émotions produites par 9 acteurs qui prononcent 2 séquences de syllabes ("hätt sandik prong nju ventsie" et "fi gött leich jean kill gos terr") ont été sélectionnées. Quarante-quatre paramètres acoustiques ont été extraits pour chacune des 144 expressions produites. Les paramètres acoustiques ont été standardisés séparément pour chaque locuteur, dans le but de neutraliser une partie des différences attribuables aux locuteurs. Une analyse factorielle a permis de réduire à 9 le nombre de ces paramètres qui étaient en partie fortement corrélés. Deux paramètres sélectionnés par l'analyse factorielle ont été exclus de la suite des analyses en raison de leur absence de relation avec les émotions exprimées. Un paramètre – l'intensité acoustique moyenne (int.moy) – a été ajouté en raison de son importance théorique dans la communication vocale des émotions. Au total 8 paramètres acoustiques ont donc été examinés: l'*intensité moyenne* (int.moy), l'étendue de l'intensité (int.étdu), le minimum de F0 (F0.min), l'étendue de F0 (F0.étdu), la durée totale (dur.tot), la durée des segments voisée relativement aux segments de parole (dur.v/art), la proportion d'énergie dans les segments voisés en dessous de 1000 Hz (v.0-1k), la proportion d'énergie dans les segments voisés comprise entre 600 et 800 Hz (v.600-800).

Des jugements relatifs à 8 caractéristiques vocales perçues ont été obtenus. Quinze à seize jugements relatifs à la *hauteur*, à l'*intensité*, à l'*intonation*, à la *rapidité*, à la qualité de l'*articulation*, à l'*instabilité* de la parole, à la qualité *rauque* et à la qualité *perçante* de la voix ont été obtenus pour chacune des 144 expressions émotionnelles. Des jugements d'intensité pour 4 émotions ont également été obtenus. Quatorze jugements relatifs à l'intensité perçue de la colère, de la peur, de la tristesse et de la joie ont été obtenus pour chacune des 144 expressions émotionnelles. Les jugements moyens (relatifs aux caractéristiques vocales perçues et aux émotions perçues) pour chaque expression seront utilisés dans la section qui suit.

L'émotion exprimée dans les 144 expressions vocales examinées est conçue comme correspondant à 4 familles (ou types) d'émotions: la colère, la joie, la peur et la tristesse, croisées avec deux niveaux d'activation. Les 8 catégories d'émotion exprimées correspondent donc à 4 types d'émotions avec un niveau d'activation faible – la colère froide, la joie calme, l'anxiété et la tristesse – et à 4 types d'émotions avec un niveau d'activation fort – la colère chaude, la joie intense, la peur panique et le désespoir. Dans les chapitres précédents, l'émotion exprimée a été traitée en premier lieu comme une variable catégorielle à 8 niveaux; l'influence des 2 niveaux d'activation postulés a été

considérée dans un deuxième temps de manière à estimer l'importance de l'effet de l'activation sur les résultats obtenus pour différentes émotions exprimées. Dans la section qui suit, l'émotion exprimée est abordée sous un angle différent. Pour chaque type d'émotion – c'est-à-dire la colère, la peur, la joie et la tristesse – une variable dichotomique a été créée de manière à représenter la présence versus l'absence du type d'émotion considéré. En conséquence, pour la nouvelle variable 'colère exprimée', les 36 expressions de colère froide et de colère chaude ont reçu la valeur 1, les 108 autres expressions ont reçu la valeur 0; pour la nouvelle variable 'peur exprimée', les expressions d'anxiété et de peur panique ont reçu la valeur 1, les autres expressions ont reçu la valeur 0; pour la nouvelle variable 'tristesse exprimée', les expressions de tristesse et de désespoir ont reçu la valeur 1, les autres expressions ont reçu la valeur 0; et pour la nouvelle variable 'joie exprimée', les expressions de joie calme et de joie intense ont reçu la valeur 1, les autres expressions ont reçu la valeur 0.

Dans cette configuration, l'influence du niveau d'activation associé à l'émotion exprimé est contrôlé. Pour chaque variable dichotomique, la moitié des expressions qui correspondent à la présence d'une émotion donnée possède un niveau d'activation postulé faible et l'autre moitié un niveau d'activation postulé fort; ces expressions sont à chaque fois opposées aux autres expressions dont la moitié comprend également un niveau d'activation défini comme faible et l'autre moitié un niveau d'activation défini comme fort.

5.3.1 Lens Model Equation (proposition de Juslin, 1998)

Le modèle utilisé par Juslin (1998) a été appliqué aux données décrites ci-dessus. Les figures 4 à 11 présentent l'ensemble des résultats obtenus. Les quatre premières figures représentent les modèles obtenus en utilisant les 8 paramètres acoustiques qui ont été sélectionnés dans la section consacrée aux résultats obtenus sur le plan acoustiques, pour la communication de la colère (figure 4), de la joie (figure 5), de la peur (figure 6) et de la tristesse (figure 7). Les quatre figures suivantes représentent les modèles obtenus en utilisant les 8 caractéristiques vocales perçues qui ont été décrites dans la section consacrée aux caractéristiques vocales perçues, pour la communication de la colère (figure 8), de la joie (figure 9), de la peur (figure 10) et de la tristesse (figure 11).

Dans ces figures, l'émotion perçue, représentée sur la droite des modèles, correspond aux jugements moyens d'intensité émotionnelle perçue décrits dans la section consacrée aux émotions attribuées aux expressions vocales (intensité de colère pour les figures 4 et 8; intensité de joie pour les figures 5 et 9; intensité de peur pour les figures 6 et 10; et intensité de tristesse pour les figures 7 et 11).

Ces figures représentent les paramètres qui constituent la *composante linéaire* de la Lens Model Equation (LME), c'est-à-dire la *validité écologique* (R_e, la corrélation multiple entre les paramètres

acoustiques ou les caractéristiques vocales perçues et l'émotion exprimée), la validité fonctionnelle (R_s, la corrélation multiple entre les paramètres acoustiques ou les caractéristiques vocales perçues et l'émotion perçue) et le *matching* (G, la corrélation entre les variables prédites dérivées à partir de la régression des paramètres acoustiques ou des caractéristiques vocales perçues sur l'émotion exprimée et sur l'émotion perçue). Les figures représentent également l'*achievement* (r_a, la corrélation entre l'émotion exprimée et l'émotion perçue). En outre, les coefficients de régression standardisés (betas) obtenus pour la régression des paramètres acoustiques ou des caractéristiques vocales perçues sur l'émotion exprimée (à droite) et respectivement sur l'émotion perçue (à gauche) ont été ajoutés sur ces figures. Les coefficients significativement différents de zéro ($p<.05$) sont signalés par un astérisque. En raison de la forte colinéarité qui existe entre les prédicteurs, ces coefficients sont relativement instables. Ils indiquent néanmoins quelles caractéristiques acoustiques ou perçues ont contribué aux prédictions résumées par les corrélations multiples. Ainsi, on observe sur la figure 4 que l'intensité moyenne, le minimum de F0 et la proportion d'énergie en dessous de 1000Hz dans les segments voisés contribuent significativement à la prédiction de la colère exprimée et de la colère perçue (pour laquelle l'énergie comprise entre 600 et 800 Hz ajoute également une contribution significative). En revanche, pour la joie (figure 5) seule l'étendue de la F0 contribue significativement à la prédiction de la joie perçue.

Figure 4: LME pour la communication de la **colère** avec 8 paramètres **acoustiques**

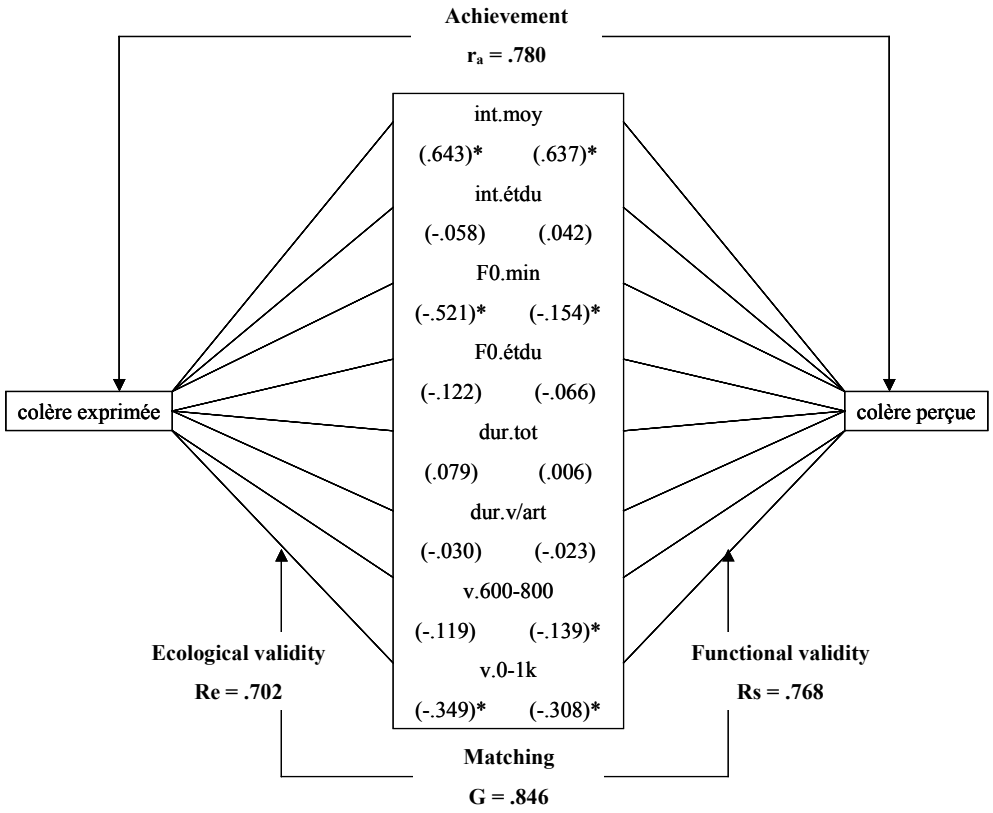


Figure 5: LME pour la communication de la **joie** avec 8 paramètres **acoustiques**

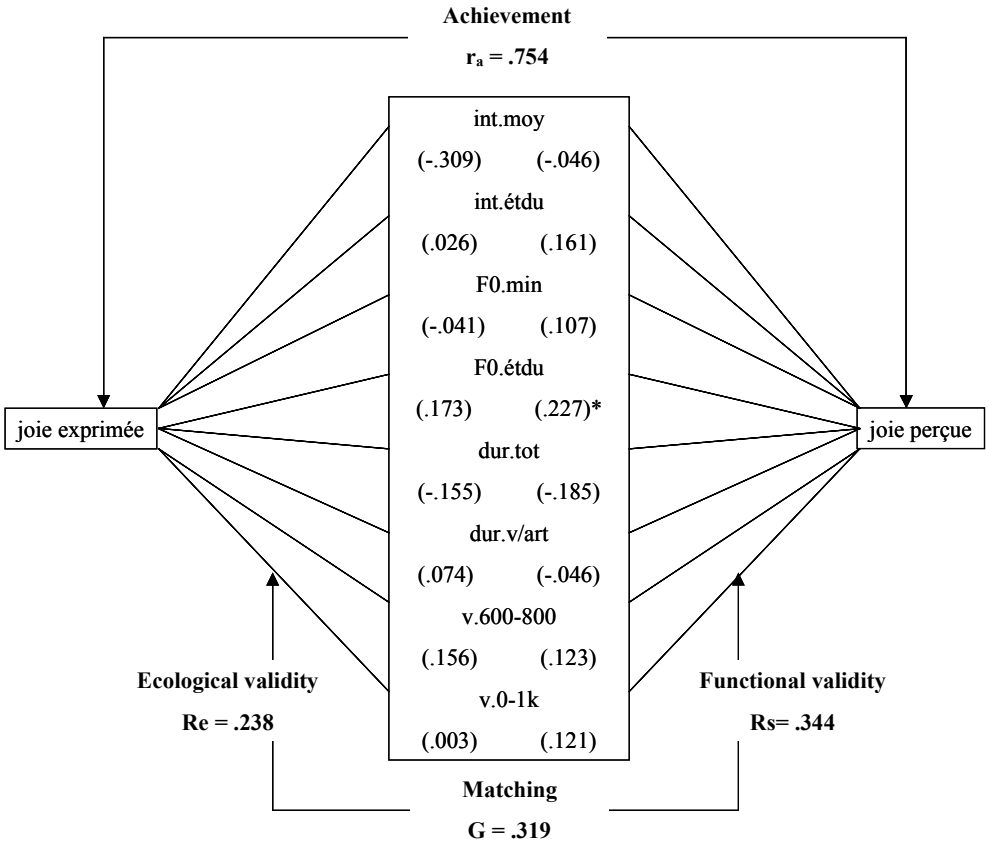


Figure 6: LME pour la communication de la **peur** avec 8 paramètres **acoustiques**

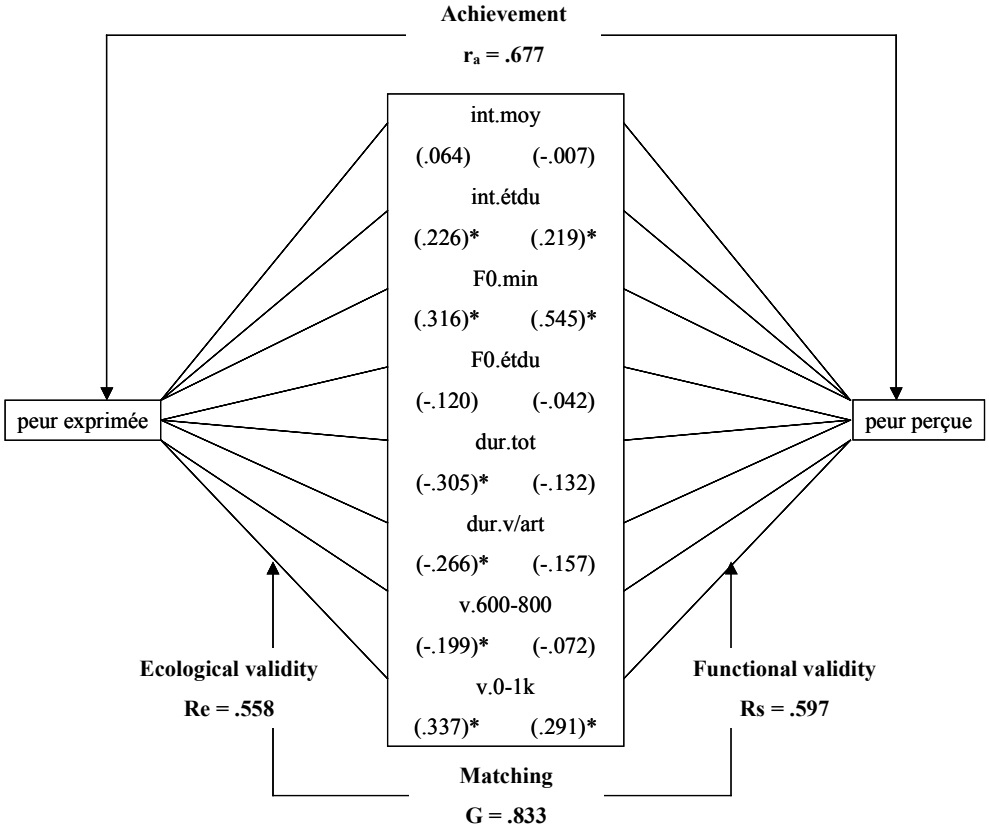


Figure 7: LME pour la communication de la **tristesse** avec 8 paramètres **acoustiques**

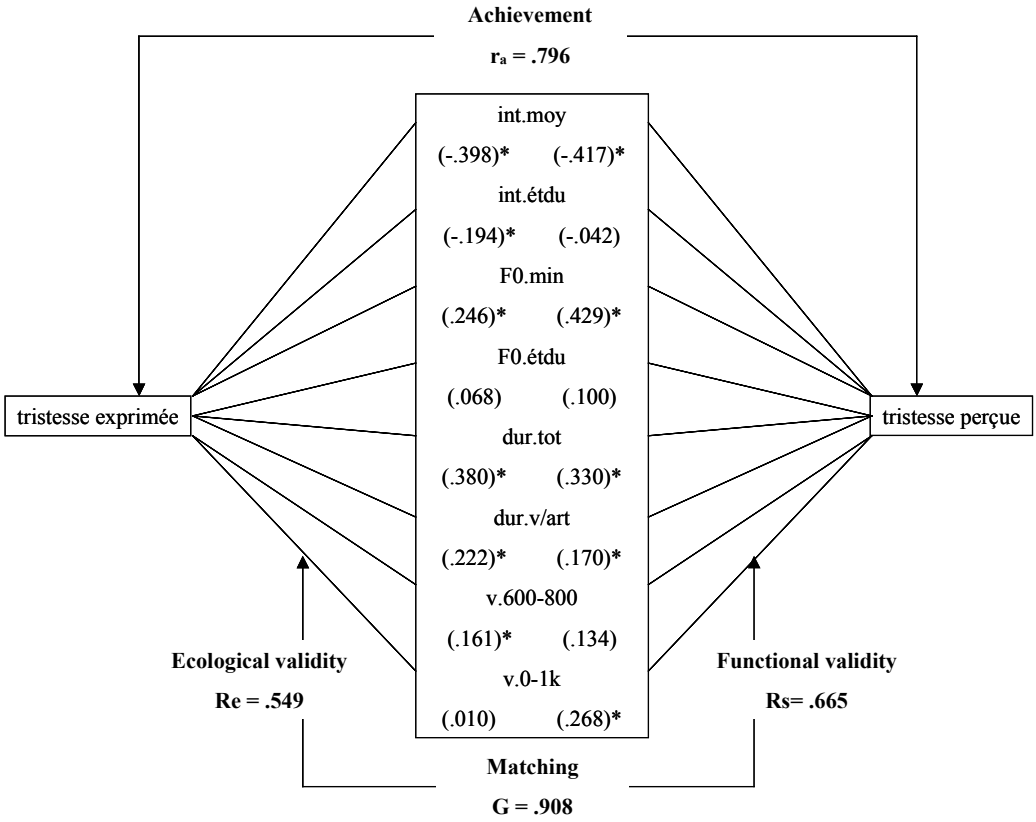


Figure 8: LME pour la communication de la **colère** avec 8 caractéristiques vocales **perçues**

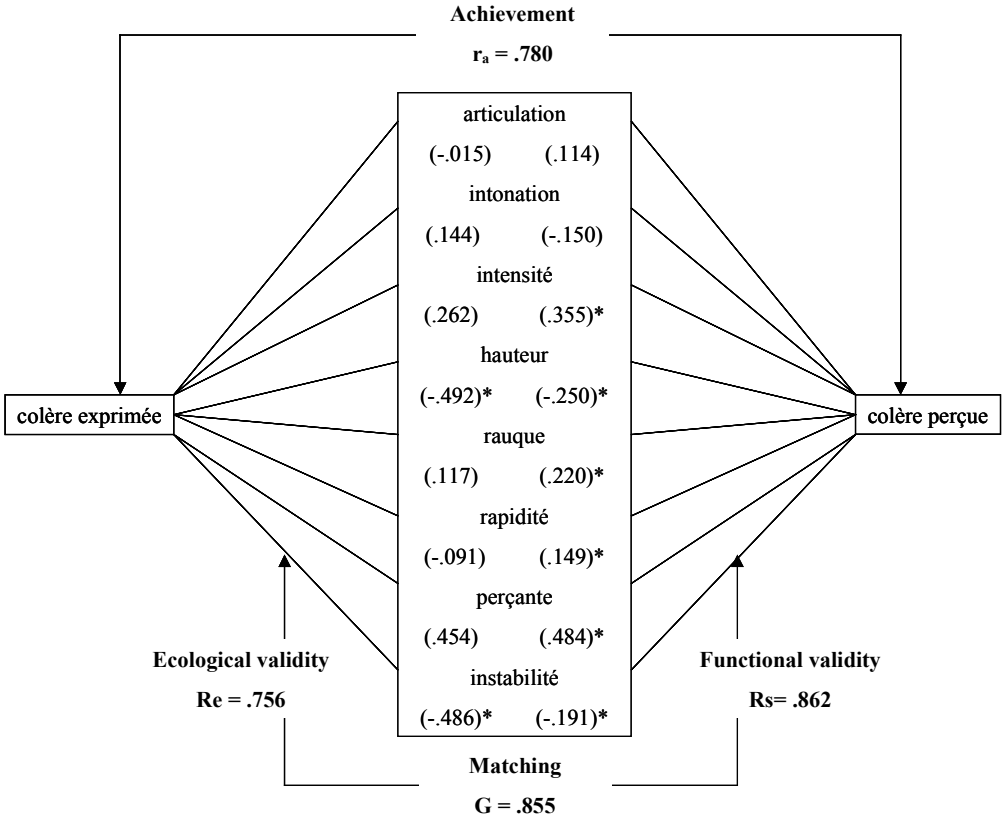


Figure 9: LME pour la communication de la **joie** avec 8 caractéristiques vocales **perçues**

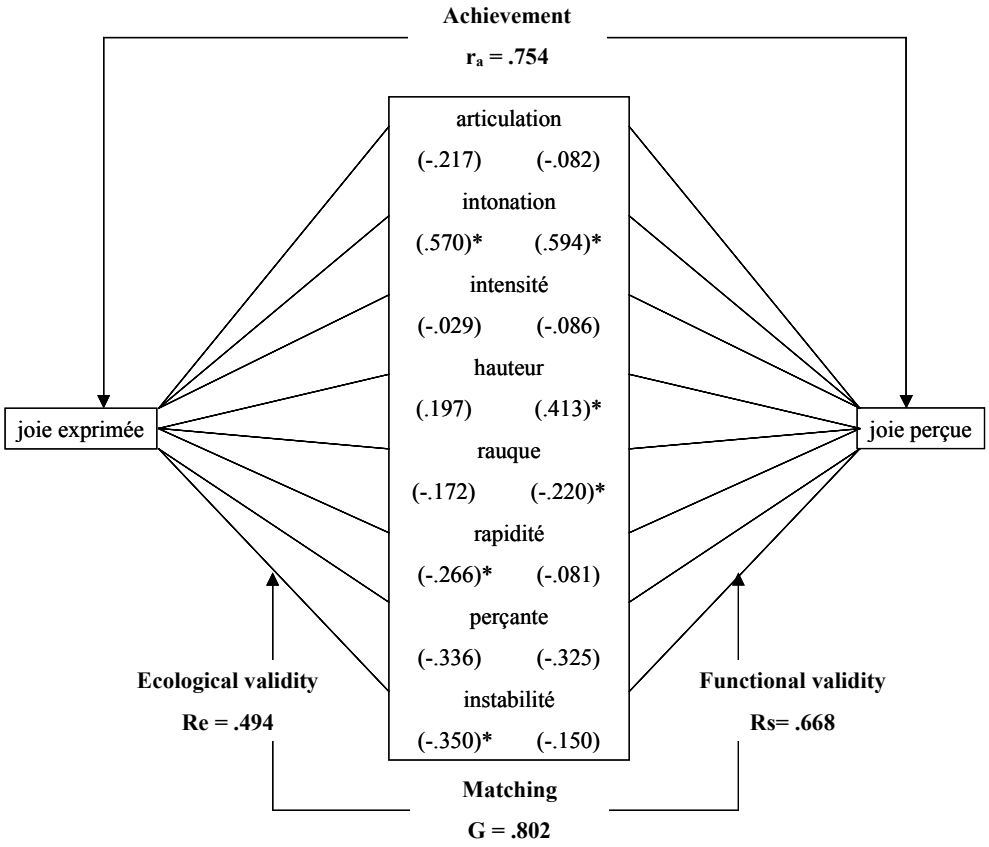


Figure 10: LME pour la communication de la **peur** avec 8 caractéristiques vocales **perçues**

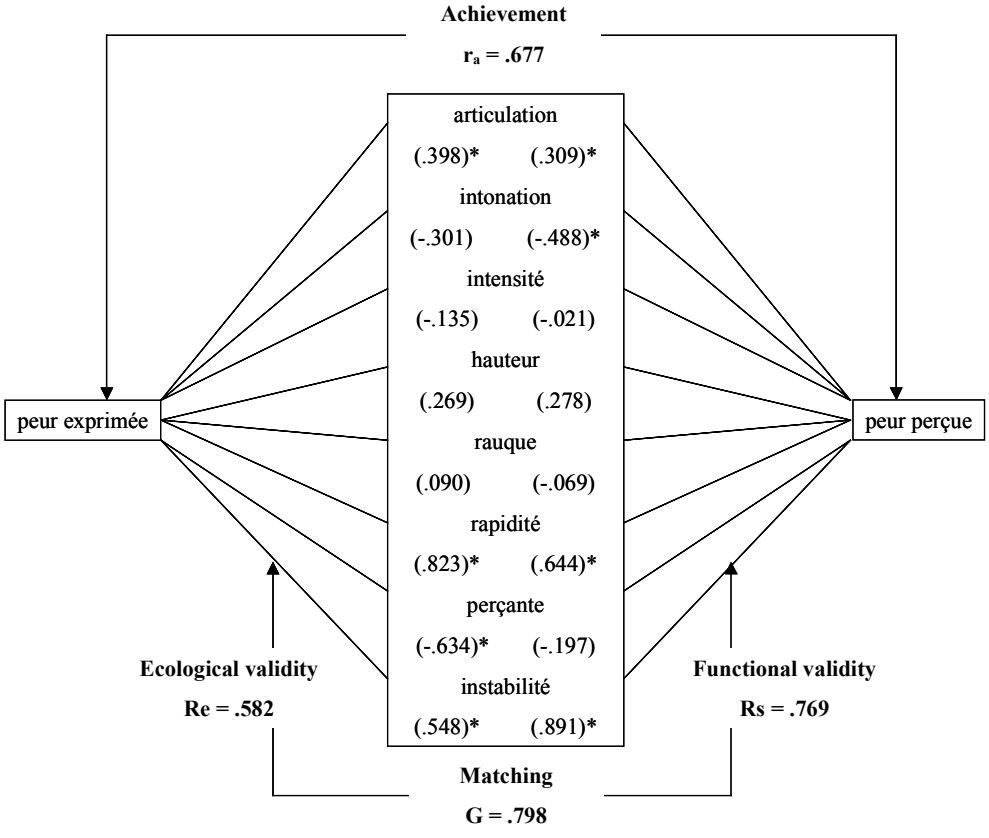
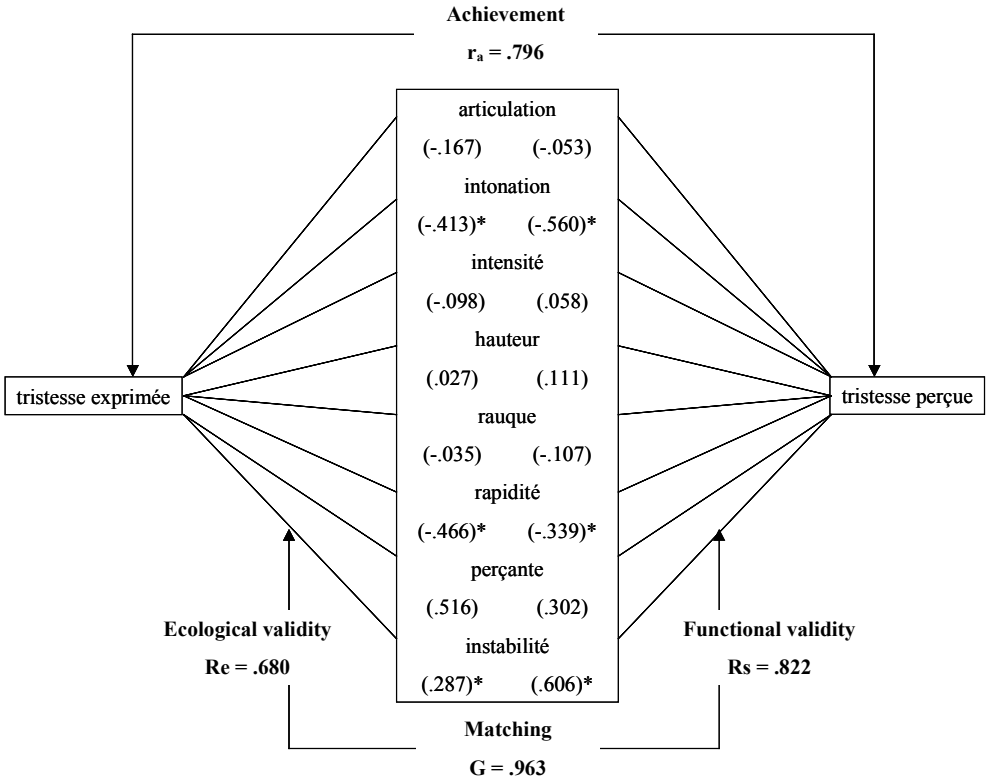


Figure 11: LME pour la communication de la **tristesse** avec 8 caractéristiques vocales **perçues**



Les résultats pour la LME complète (c'est-à-dire pour la composante linéaire et pour la composante non-modélisée) sont résumés dans le tableau 2. Les quatre premières lignes de ce tableau reprennent les résultats déjà présentés dans les figures 4 à 11 – c'est-à-dire la valeur qui correspond à l'*achievement* et les trois paramètres de la composante linéaire du modèle – pour les quatre émotions communiquées et pour les modèles dérivés des paramètres acoustiques, ainsi que pour les modèles dérivés des caractéristiques vocales perçues. La cinquième ligne représente la valeur correspondant à la composante linéaire de chaque modèle. Les trois lignes suivantes représentent les trois paramètres qui définissent la composante non-modélisée et la neuvième ligne représente la valeur totale correspondant à la composante non-modélisée. Finalement, la dernière ligne représente la proportion de l'*achievement* (corrélation entre l'émotion exprimée et l'émotion perçue) qui est modélisée par les paramètres acoustiques ou par les caractéristiques vocales perçues pour chaque émotion communiquée.

Les figures 4 à 11 et le tableau 2 indiquent que le coefficient d'*achievement* varie relativement faiblement pour les quatre types d'émotions (de .68 pour la peur à .80 pour la tristesse). Ils révèlent également que les paramètres acoustiques et les caractéristiques vocales perçues ne modélisent pas la même proportion de l'*achievement* pour les quatre types d'émotions communiquées. Le "meilleur" modèle est obtenu pour la colère, 58% de l'*achievement* pour cette émotion est modélisé par les paramètres acoustiques et 71% par les caractéristiques vocales perçues. Le modèle le "moins

bon" est obtenu pour la joie, avec seulement 3% de l'*achievement* modélisé par l'intermédiaire des paramètres acoustiques et 35% de l'*achievement* modélisé par les caractéristiques vocales perçues. L'*achievement* pour la peur et la tristesse semble un peu moins bien modélisé que l'*achievement* pour la colère; les modèles pour ces deux émotions sont toutefois nettement meilleurs que pour la joie.

En considérant séparément les quatre types d'émotions communiquées, on remarque que les modèles obtenus à l'aide des caractéristiques vocales perçues expliquent systématiquement une proportion plus importante de l'*achievement* que les modèles obtenus à l'aide des paramètres acoustiques. Cette différence ne peut être attribuée uniquement à une meilleure modélisation de l'émotion perçue par les caractéristiques perçues. Les coefficients R_s et R_e sont systématiquement plus élevés pour les modèles qui utilisent les caractéristiques vocales perçues. Relativement aux paramètres acoustiques utilisés, les caractéristiques vocales semblent donc réussir une modélisation légèrement meilleure non seulement de l'émotion perçue, mais également de l'émotion exprimée.

Tableau 2: LME pour la communication de la colère, de la joie, de la peur et de la tristesse, modèles obtenus séparément pour 8 paramètres acoustiques et pour 8 caractéristiques vocales perçues

	colère		joie		peur		tristesse	
coefficients	acoust.	perçues	acoust.	perçues	acoust.	perçues	acoust.	perçues
r_a	**0.780**	**0.780**	**0.754**	**0.754**	**0.677**	**0.677**	**0.796**	**0.796**
G	0.846	0.855	0.319	0.802	0.833	0.798	0.908	0.963
R_e	0.702	0.756	0.238	0.494	0.558	0.582	0.549	0.680
R_s	0.768	0.862	0.344	0.668	0.597	0.769	0.665	0.822
$G R_e R_s$	**0.456**	**0.557**	**0.026**	**0.265**	**0.277**	**0.357**	**0.331**	**0.538**
C	0.711	0.672	0.798	0.756	0.600	0.615	0.745	0.618
$\sqrt{1-R_e^2}$	0.712	0.655	0.971	0.869	0.830	0.813	0.836	0.733
$\sqrt{1-R_s^2}$	0.640	0.507	0.939	0.744	0.802	0.639	0.747	0.569
$C\sqrt{1-R_e^2}\sqrt{1-R_s^2}$	**0.324**	**0.223**	**0.728**	**0.489**	**0.399**	**0.320**	**0.465**	**0.258**
$G R_e R_s / r_a$	0.58	0.71	0.03	0.35	0.41	0.53	0.42	0.68

Pour les huit modèles présentés, l'encodage (représenté par la corrélation multiple R_e) est légèrement moins bien modélisé que le décodage (représenté par la corrélation multiple R_s). Les différences les plus importantes sont observées pour les modèles basés sur les caractéristiques vocales perçues pour la joie (proportion de variance expliquée pour la joie exprimée: $R^2_e = 0.244$, proportion de variance expliquée pour la joie perçue: $R^2_s = 0.446$, $R^2_s - R^2_e = 0.202$), la peur ($R^2_e = 0.339$, $R^2_s = 0.591$, $R^2_s - R^2_e = 0.252$) et pour la tristesse ($R^2_e = 0.462$, $R^2_s = 0.676$, $R^2_s - R^2_e =$

0.214). Cette différence est plus faible pour les modèles basés sur les paramètres acoustiques qui réalisent donc une prédiction légèrement plus "équilibrée" du décodage et de l'encodage.

La composante non-modélisée est dans l'ensemble assez importante. Le "meilleur" modèle (basé sur les caractéristiques vocales perçues pour la communication de la colère) parvient à médiatiser un peu plus des deux tiers (71%) de l'*achievement*. La composante non-modélisée est particulièrement importante pour la communication de la joie. Le modèle basé sur les paramètres acoustiques pour cette émotion ne parvient pas à rendre compte de la communication; le modèle basé sur les caractéristiques vocales perçue parvient en revanche à rendre d'un plus du tiers (35%) de l'*achievement*. La prédiction de la joie exprimée (R_e^2 = 0.057, soit 6% de variance expliquée) et de la joie perçue (R_s^2 = 0.118, soit 12% de variance expliquée) étant très faibles pour le modèle basé sur les paramètres acoustiques, le *matching* entre l'encodage et le décodage (G^2 = 0.102, soit 10% de variance partagée entre la prédiction de la joie exprimée et de la joie perçue) est également très faible. En revanche l'*achievement* est aussi élevé pour la communication de la joie que pour les autres émotions. Ceci indique que les caractéristiques vocales (acoustiques ou perçues) qui peuvent être utilisées pour médiatiser la communication de certaines émotions (ici la colère, la peur et la tristesse) ne permettent pas nécessairement de modéliser la communication d'autres émotions (dans ce cas la joie).

Dans les modèles présentés ci-dessus, le rôle de l'activation émotionnelle est contrôlé en utilisant un nombre identique d'expressions fortement activées et d'expressions faiblement activées pour chaque émotion exprimée. Techniquement, il est donc aisé de définir une nouvelle variable dichotomique 'activation exprimée' en procédant comme pour les variables dichotomiques 'émotion (colère, joie, peur, tristesse) exprimée', c'est-à-dire en attribuant une valeur 0 aux émotions faiblement activées et une valeur 1 aux émotions fortement activées. En revanche, nous ne disposons pas de jugements pour *l'activation perçue*. Afin de présenter malgré tout une évaluation du rôle de l'activation sous-jacente aux émotions exprimées dans le cadre de la LME, une estimation de l'*intensité émotionnelle perçue* a été définie pour chaque expression vocale; elle correspond à la moyenne des jugements d'intensité émotionnelle obtenus pour la colère, joie, la peur et la tristesse. La relation entre l'activation exprimée et l'intensité perçue a été ensuite modélisée de la même manière que les relations entre les émotions exprimées et les émotions perçues.

Les figures 12 et 13, ainsi que le tableau 3 présentent les valeurs correspondant aux deux modèles obtenus pour la relation entre l'activation exprimée et l'intensité perçue, basés respectivement sur les paramètres acoustiques et sur les caractéristiques vocales perçues. Il existe au moins trois différences importantes entre ces modèles et ceux obtenus pour la communication des émotions: (1) L'aspect encodé (l'activation associée à l'émotion exprimée) est lié, mais n'est pas conceptuellement équivalent à l'aspect décodé (intensité émotionnelle moyenne perçue). (2) La distribution de la variable dichotomique 'activation exprimée' est différente de la distribution des variables dichotomiques 'colère/joie/peur/tristesse exprimée'. La moitié des expressions sont fortement activées, respectivement faiblement activées, alors qu'un quart des expressions seulement correspondent à chaque type d'émotion. (3) La variance de l'intensité moyenne perçue est comparativement faible. Chaque expression peut obtenir un ou plusieurs jugements émotionnels élevés, mais les jugements moyens (pour les 4 intensités émotionnelles n'atteignent au maximum que des valeurs inférieures à 6 (sur une échelle qui varie théoriquement de 0 à 10).

Figure 12: LME pour la relation entre **activation et intensité perçue** avec 8 paramètres **acoustiques**

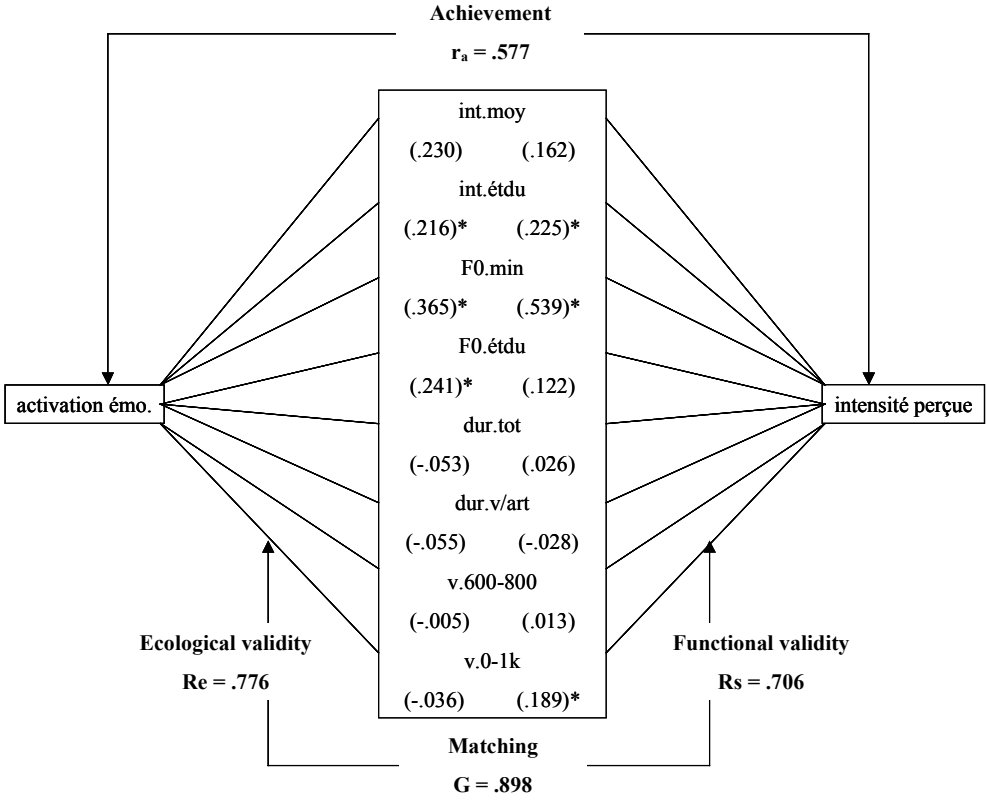


Figure 13: LME pour la relation entre **activation et intensité perçue** avec 8 caractéristiques vocales **perçues**

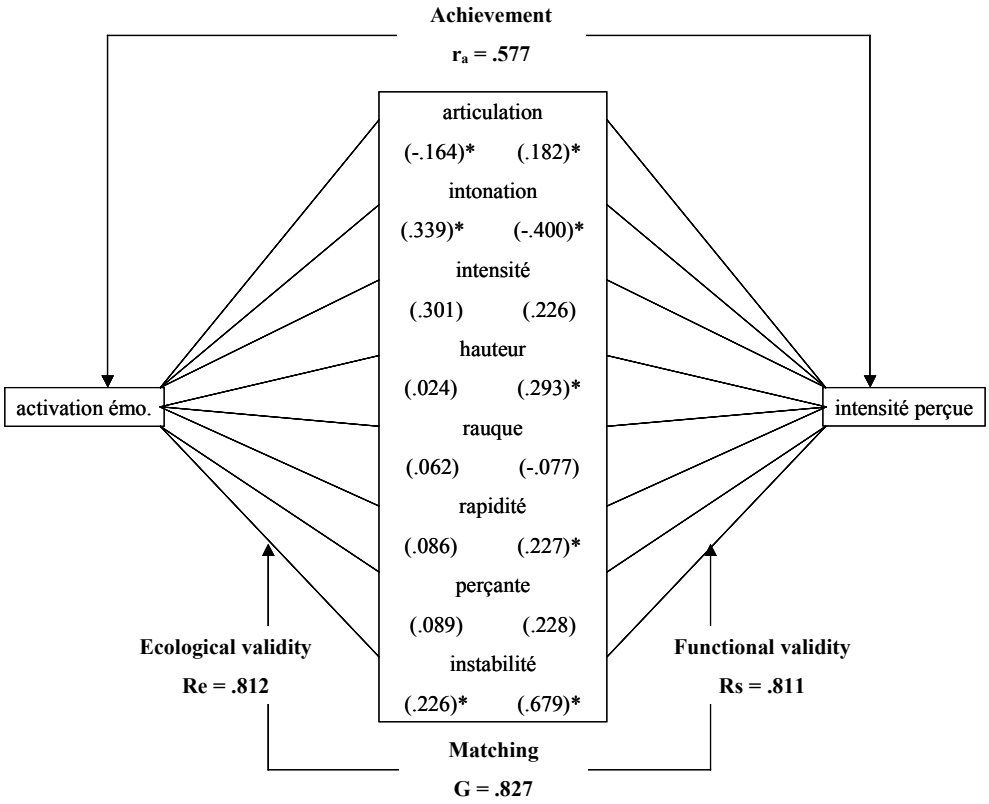


Tableau 3: LME pour la relation entre activation et intensité perçue, modèles obtenus séparément pour 8 paramètres acoustiques et pour 8 caractéristiques vocales perçues

	activation/intensité			activation/intensité	
coefficients	acoust.	perçues	coefficients	acoust.	perçues
r_a	**0.577**	**0.577**	C	0.190	0.095
G	0.898	0.827	$\sqrt{1-R_e^2}$	0.631	0.584
R_e	0.776	0.812	$\sqrt{1-R_s^2}$	0.708	0.585
R_s	0.706	0.811	$C\sqrt{1-R_e^2}\sqrt{1-R_s^2}$	**0.085**	**0.032**
$G R_e R_s$	**0.492**	**0.545**	$G R_e R_s / r_a$	0.85	0.94

Logiquement, au regard de la différence conceptuelle qui existe entre l'activation et l'intensité, la corrélation entre l'activation exprimée est l'intensité moyenne perçue est plus faible que les corrélations entre les émotions exprimées et les émotions perçue. Un lien non-négligeable existe toutefois entre ces deux variables ($r^2 = 0.333$, soit un tiers de variance partagée). Le modèle basé sur les paramètres acoustiques et, surtout, le modèle basé sur les caractéristiques vocales perçues parviennent à représenter une très grande partie de ce lien (respectivement 85% et 94% de la

corrélation). Si l'on compare ces deux modèles avec les modèles obtenus pour la communication de la colère – qui correspond à l'émotion la mieux modélisée – on remarque que la prédiction de l'activation exprimée par les caractéristiques vocale (acoustiques et perçues) est légèrement meilleure que la prédiction de la colère exprimée. En revanche, la prédiction de l'intensité perçue (moyenne des 4 intensités émotionnelles jugées) n'est pas meilleure que la prédiction de l'intensité de colère perçue. Ces résultats soulignent une nouvelle fois la grande capacité des paramètres acoustiques à "capturer" le niveau d'activation associé aux émotions exprimées. Le coefficient G est aussi important pour le modèle activation-intensité que pour les modèles qui représentent la communication des émotions; les paramètres acoustiques et les caractéristiques vocales perçues sont utilisés de manière très similaire pour la prédiction de l'activation et pour la prédiction de l'intensité moyenne perçue. La "supériorité" des modèles obtenus pour la relation entre l'activation exprimée et l'intensité perçue réside surtout dans la valeur beaucoup plus faible du coefficient C pour ces modèles relativement aux modèles présentés pour la communication des émotions. Les faibles valeurs du coefficient C pour les modèles présentés dans le tableau 3 signalent que la variance résiduelle de l'activation et la variance résiduelle de l'intensité perçue correspondent probablement à des composantes d'erreur non-systématique.

D'autres manières de décomposer l'*achievement* ont été proposées par différents auteurs qui ont développé et/ou utilisé diverses extensions de la Lens Model Equation (v. Stewart, 2001, pour une revue des applications et de extensions proposées pour la LME). Certaines extensions de la LME permettent d'inclure différents types de variables dans la décomposition de l'*achievement*. Un exemple d'extension de la LME qui permet d'inclure à la fois les paramètres acoustiques et les propriétés vocales dans un seul modèle pour la communication de chaque émotion est présenté en annexe (v. annexe D1 pour une description de cette proposition et son application aux données). Ces modèles n'ajoutent toutefois que peu d'information supplémentaire aux modèles présentés ci-dessus. Une perspective réellement différente est en revanche obtenue en utilisant la décomposition de l'*achievement* proposée par Scherer (1978).

5.3.2 Path Analysis (proposition de Scherer, 1978)

Le modèle statistique proposé par Scherer utilise simultanément les variables acoustiques et les variables représentant les caractéristiques vocales perçues. Nous avons mesuré et sélectionné au total 18 variables: l'émotion exprimée et l'émotion perçue, 8 paramètres acoustiques et 8 caractéristiques vocales perçues. Au regard des motifs suivants, ce nombre de variables est trop important pour être utilisé dans un même modèle statistique: (1) Pour l'utilisation de la régression, le nombre de variable ne doit pas être disproportionné par rapport au nombre de cas examinés. Les

données analysées comportent 144 cas, ce qui correspondrait à une proportion de seulement 8 cas par variable au cas où un modèle inclurait 18 variables; or il est conseillé que la relation entre le nombre de variables et le nombre de cas analysés soit plutôt de l'ordre de 20 cas par variable. (2) La colinéarité entre ces variables est assez importante; or plus le nombre de variables fortement colinéaires est augmenté, plus les résultats de la régression sont instables et difficiles à interpréter. (3) Au-delà d'un certain nombre de variables, la lisibilité du modèle proposé par Scherer est compromise.

En conséquence, nous avons choisi de limiter le nombre de variables acoustiques ainsi que le nombre de variables représentant les caractéristiques vocales perçues à trois par modèle, de manière à construire pour chaque émotion communiquée un modèle incluant au total huit variables. Une décision a été prise de sélectionner les paramètres acoustiques et les caractéristiques vocales perçues les plus susceptibles de médiatiser la communication de chaque type d'émotion (colère, peur, joie et tristesse), c'est-à-dire d'effectuer une nouvelle sélection pour chaque émotion communiquée. Des régressions "progressives" (stepwise) ont été effectuées afin de sélectionner ces variables. Pour chaque émotion exprimée, les trois paramètres acoustiques qui sont parvenus à expliquer la plus forte proportion de la variance de l'émotion exprimée ont été sélectionnés. Pour chaque émotion perçue, les trois caractéristiques vocales perçues qui sont parvenues à expliquer la plus forte proportion de la variance de l'émotion perçue ont été sélectionnées. Les paramètres acoustiques ont été sélectionnés relativement à l'émotion exprimée, alors que les caractéristiques vocales perçues ont été sélectionnées relativement à l'émotion perçue, car le modèle proposé par Scherer stipule une plus grande proximité entre les caractéristiques vocales perçues et l'émotion perçue (relativement à l'émotion exprimée) et, respectivement, une plus grande proximité entre les paramètres acoustiques et l'émotion exprimée (relativement à l'émotion perçue).

Les résultats des régressions (stepwise) destinées à sélectionner les paramètres acoustiques et les caractéristiques vocales perçues sont présentés dans le tableau 4. La corrélation multiple et l'erreur standard sont présentés pour les modèles incluant les trois premières variables entrées dans chaque analyse.

Pour la tristesse exprimée, seuls deux paramètres acoustiques ajoutent une contribution indépendante significative à la prédiction de l'émotion exprimée. Pour la joie exprimée, aucun paramètre acoustique ne contribue significativement à la prédiction de l'émotion exprimée. Pour la joie perçue, seuls deux caractéristiques vocales perçues ajoutent une contribution indépendante significative à la prédiction de l'émotion perçue. Afin de présenter des modèles basés sur un nombre identique de variables pour les 4 émotions, des variables supplémentaires ont dû être sélectionnées.

Pour la tristesse exprimée, la variable acoustique qui présente la plus forte corrélation avec cette émotion exprimée – soit l'intensité acoustique moyenne (int.moy, R = -0.309) – a été sélectionnée. De même pour la joie perçue, la caractéristique vocale perçue qui présente la plus forte corrélation avec cette émotion perçue – soit la hauteur perçue (R = 0.472) – a été sélectionnée. En revanche, aucun paramètre acoustique n'étant significativement corrélé à la joie exprimée, la sélection pour cette émotion exprimée est nécessairement plus arbitraire et n'aura pas, ou peu, d'influence sur le résultat final obtenu (la relation entre la joie exprimée et les paramètres acoustiques sera dans tous les cas nulle). Un paramètre lié à la durée (la durée totale, dur.tot), un paramètre lié à la F0 (l'étendue de F0, F0.étdu) et un paramètre lié à l'intensité (l'intensité moyenne, int.moy) ont été arbitrairement sélectionnés.

Tableau 4: Régressions (stepwise) destinées à sélectionner les 3 meilleurs prédicteurs pour l'émotion exprimée parmi les paramètres acoustiques et les 3 meilleurs prédicteurs pour l'émotion perçue parmi les caractéristiques vocales perçues.

	Paramètres acoustiques	R	R^2	R^2 ajusté	Erreur standard*
colère exprimée	v.0-1k, F0.min, int.moy	0.686	0.471	0.459	0.319
peur exprimée	F0.min, v.0-1k, dur.tot	0.446	0.199	0.182	0.393
tristesse exprimée	dur.tot, int.étdu, -	0.470	0.220	0.209	0.386
joie exprimée	- - -	-	-	-	-
	Dimensions vocales	**R**	**R^2**	**R^2 ajusté**	**Erreur standard***
colère perçue	intensité, hauteur, perçante	0.833	0.694	0.688	1.504
peur perçue	instabilité, rapidité, intonation	0.715	0.511	0.501	1.575
tristesse perçue	instabilité, intonation, rapidité	0.801	0.641	0.634	1.428
joie perçue	intonation, rauque, -	0.633	0.400	0.392	1.703

*L'erreur standard qui correspond à la racine carrée de la moyenne des carrés des résidus de la régression est une fonction de la variance de la variable prédite. L'erreur standard est plus élevée pour l'émotion perçue que pour l'émotion exprimée car la variance de l'émotion perçue qui est évaluée sur une échelle continue (de 0 à 10) est plus importante que la variance de l'émotion exprimée qui correspond à une variable dichotomique (codée 0-1).

Le tableau 5 présente les résultats obtenus pour la régression hiérarchique sur chaque émotion perçue des caractéristiques vocales perçues sélectionnées, puis des paramètres acoustiques sélectionnés et finalement de l'émotion exprimée. Ce tableau permet d'observer dans quelle mesure l'ajout des paramètres acoustiques aux caractéristiques vocales perçues, dans un premier temps, puis l'ajout de l'émotion exprimée permet d'accroître la part de variance expliquée (R^2) de l'émotion perçue. La signification statistique (sig.) de l'augmentation de la corrélation multiple pour des modèles successifs est indiqué dans la dernière colonne du tableau 5.

Les résultats présentés dans le tableau 5 indiquent que – pour la colère, la tristesse et la joie – l'ajout des paramètres acoustiques n'augmente pas significativement la taille de la corrélation multiple relativement aux modèles qui utilisent uniquement les caractéristiques vocales perçues. Pour la peur, en revanche, la part de variance expliquée est significativement augmentée par l'introduction

des paramètres acoustiques. Pour les quatre émotions, l'ajout de la variable dichotomique correspondant à l'émotion exprimée augmente significativement la part de variance expliquée par les modèles.

Tableau 5: Régressions hiérarchiques de 3 caractéristiques vocales perçues, puis de 3 paramètres acoustiques, puis de l'émotion exprimée sur les émotions perçues

	Variables acoustiques et dimensions vocales	R	R^2	F change	Sig. F change
colère perçue	perçante, hauteur, intensité	0.833	0.694	105.97	.000
	perçante, hauteur, intensité, F0.min, v.0-1k, int.moy	0.840	0.706	1.76	.158
	perçante, hauteur, intensité, F0.min, v.0-1k, int.moy, colère exprimée	0.918	0.842	118.07	.000
peur perçue	instabilité, intonation, rapidité	0.715	0.511	48.80	.000
	instabilité, intonation, rapidité, F0.min, v.0-1k, dur.tot	0.773	0.598	9.85	.000
	instabilité, intonation, rapidité, F0.min, v.0-1k, dur.tot, peur exprimée	0.857	0.735	70.24	.000
tristesse perçue	instabilité, intonation, rapidité	0.801	0.641	83.48	.000
	instabilité, intonation, rapidité, int.étdu, dur.tot, int.moy	0.807	0.651	1.27	.289
	instabilité, intonation, rapidité, int.étdu, dur.tot, int.moy, tristesse exprimée	0.891	0.794	94.88	.000
joie perçue	rauque, hauteur, intonation	0.640	0.410	32.43	.000
	rauque, hauteur, intonation, dur.tot, F0.étdu, int.moy	0.649	0.421	0.91	.440
	rauque, hauteur, intonation, dur.tot, F0.étdu, int.moy, joie exprimée	0.871	0.759	190.73	.000

Les figures 15 à 18 représentent les modèles obtenus en appliquant la procédure utilisée par Scherer (1978). La figure 15 représente les valeurs obtenues pour le modèle proposé pour la communication de la colère. La figure 16 correspond au modèle proposé pour la communication de la joie. La figure 17 représente le modèle proposé pour la peur et la figure 18 représente le modèle proposé pour la tristesse. Les valeurs indiquées sur les flèches correspondent à des corrélations (pour les flèches qui relient l'émotion exprimée aux paramètres acoustiques) ou à des coefficients de régression (betas) obtenus dans des séries de régressions multiples. Les corrélations/betas significativement différents de zéro ($p<.05$) sont indiqués en gras et sont suivis d'un astérisque.

On observe sur ces figures que des valeurs significatives apparaissent aussi bien pour les "chemins" (paths) centraux – représentés par des successions de 3 flèches noires – que pour les "chemins"

(paths) périphériques – représentés par une flèche noire suivie d'une flèche verte ou par une flèche rouge suivie d'une flèche noire – qui relient l'émotion exprimée à l'émotion perçue. Les "chemins" directs – représentés par une flèche bleue – possèdent des coefficients très élevés pour les quatre émotions.

Figure 15: Path analysis pour la communication de la **colère** avec 3 caractéristiques vocales et 3 paramètres acoustiques présélectionnés.

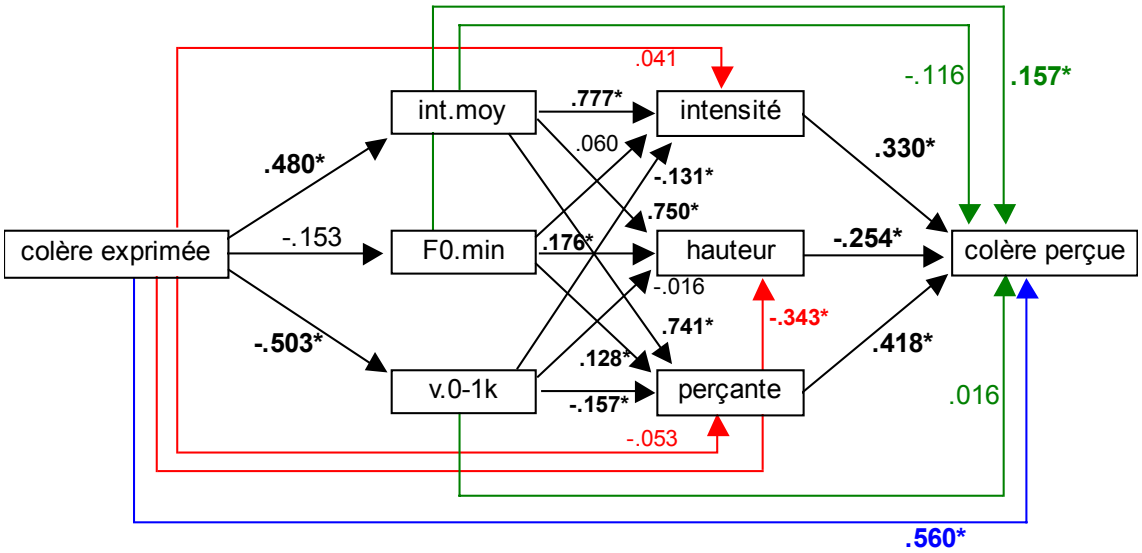
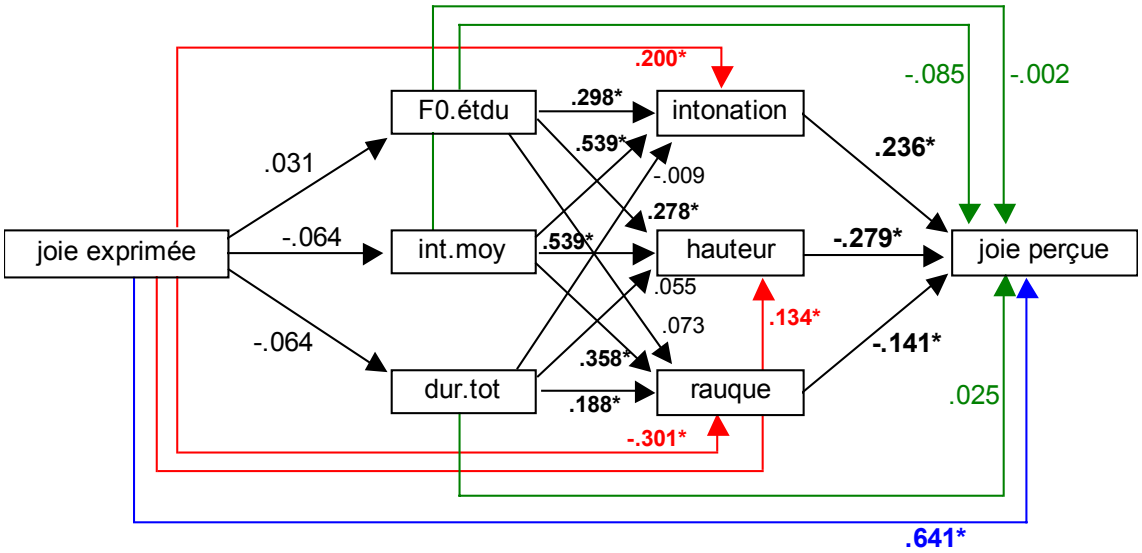


Figure 16: Path analysis pour la communication de la **joie** avec 3 caractéristiques vocales et 3 paramètres acoustiques présélectionnés.



L'interprétation des coefficients de régression représentés sur les figures 15 à 18 doit être assortie de certaines précautions. En effet, ces coefficients ne peuvent être interprétés comme indiquant des relations (positives, négatives; ou nulles) entre deux variables reliées par une flèche, indépendamment des autres variables. Les coefficients de régression sont indicatifs de l'utilisation d'une variable dans le contexte des autres variables qui figurent dans le modèle. Deux exemples

relatifs à cette difficulté d'interprétation des coefficients de régression sont développés dans les paragraphes qui suivent.

Figure 17: Path analysis pour la communication de la **peur** avec 3 caractéristiques vocales et 3 paramètres acoustiques présélectionnés.

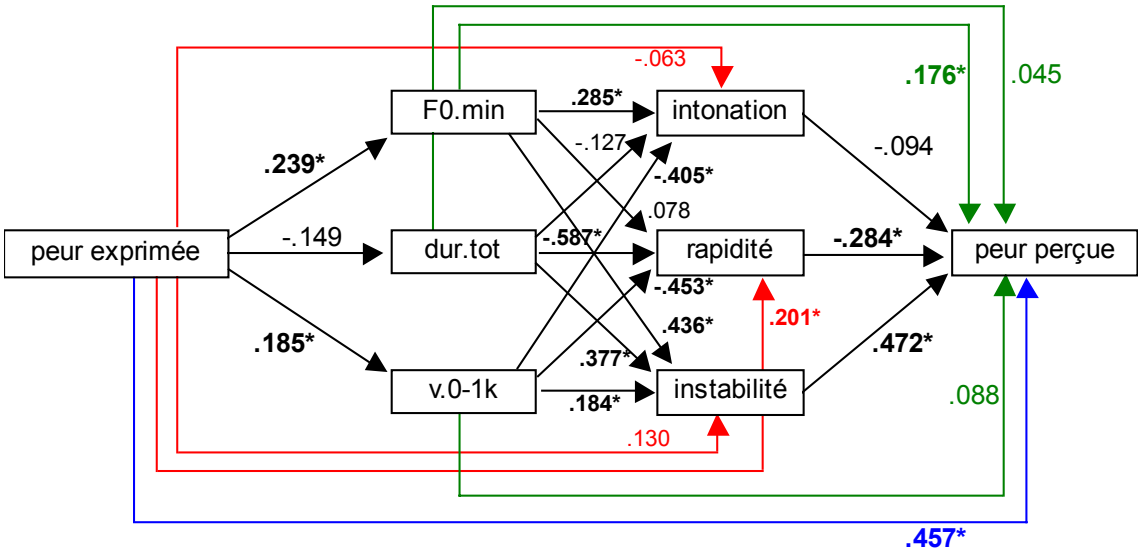
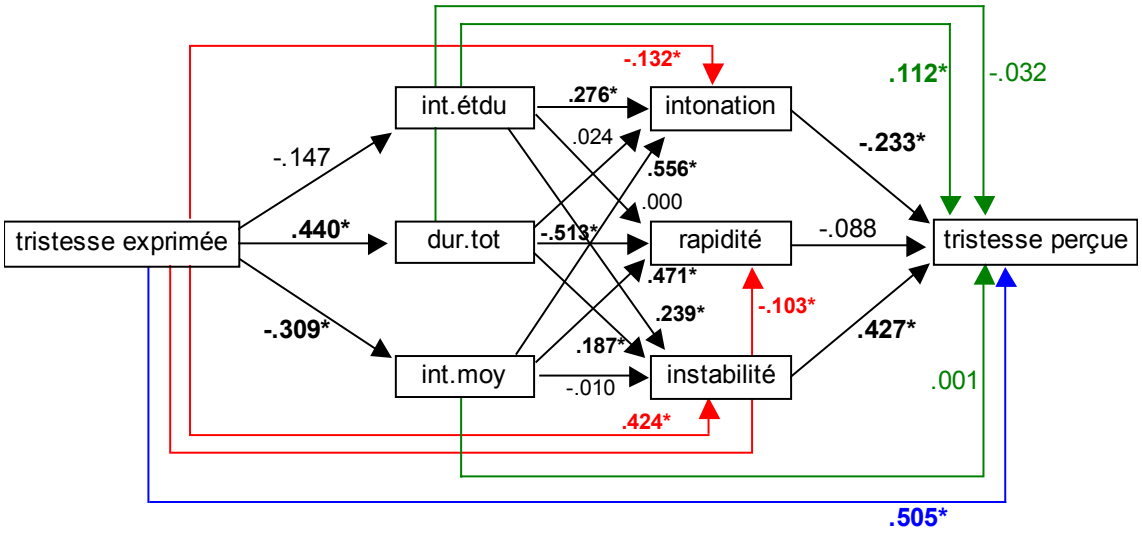


Figure 18: Path analysis pour la communication de la **tristesse** avec 3 caractéristiques vocales et 3 paramètres acoustiques présélectionnés.



Dans la figure 16 (ci-dessus) un coefficient négatif significatif relie la hauteur perçue à la joie perçue. Le lecteur peut être tenté de déduire que plus une expression est grave plus l'intensité de joie perçue est élevée. Cette interprétation est malheureusement erronée, la corrélation entre la hauteur perçue et la joie perçue est positive (et significative, $R = 0.472$, $p<.001$), les expressions perçues comme plus aiguës sont perçues comme correspondant à une intensité de joie plus forte. La hauteur perçue est utilisée dans le modèle comme un prédicteur négatif (significatif) pour une part

résiduelle de la variance de la joie perçue, d'autres variables colinéaires de la hauteur perçue "absorbent" la relation originale positive entre la hauteur perçue et la joie perçue.

Un deuxième exemple relatif à cette difficulté d'interprétation des coefficients de régression peut être observé sur la figure 18 (ci-dessus). Au vu de la littérature et aussi des résultats présentés dans les chapitres précédents, on s'attendrait à ce que la durée des expressions et respectivement la rapidité perçue jouent un rôle important dans la communication de la tristesse. Or on observe sur la figure 18 que si la tristesse exprimée est effectivement positivement corrélée avec la durée des expressions (les expressions tristes sont plus longues que les autres expressions) et que si la durée des expressions semble bien liée à la rapidité perçue (les expressions plus longues sont perçues comme moins rapides), le lien entre la rapidité perçue est la tristesse perçue est représenté par un coefficient nul. A nouveau, il est incorrect de déduire que la rapidité/durée n'est pas liée au décodage de la tristesse. La corrélation entre la rapidité perçue et la tristesse perçue est de – 0.492 ($p<.001$), ce qui indique que les expressions qui sont jugées plus lentes sont également jugées plus tristes. Dans ce cas également, la présence de colinéarité dans le modèle a pour résultat d'annuler la relation entre la rapidité perçue et la tristesse perçue dans le contexte des autres variables qui figurent dans le modèle.

La procédure utilisée ci-dessus pour la communication des émotions a été répétée pour la relation entre l'activation exprimée et l'intensité moyenne perçue. Les résultats de régressions "progressives" (stepwise) destinées à sélectionner 3 paramètres acoustiques pour la prédiction de l'activation exprimée et trois caractéristiques vocales perçues pour la prédiction de l'intensité moyenne perçue sont présentés dans le tableau 6.

Tableau 6: Régressions (stepwise) destinées à sélectionner les 3 meilleurs prédicteurs pour l'activation exprimée parmi les paramètres acoustiques et les 3 meilleurs prédicteurs pour l'intensité moyenne perçue parmi les caractéristiques vocales perçues.

	Paramètres acoustiques/ Dimensions vocales	R	R^2	R^2 ajusté	Erreur standard*
activation	int.moy, F0.min, int.étdu	0.754	0.569	0.560	0.333
intensité	hauteur, instabilité, intensité	0.778	0.605	0.596	0.603

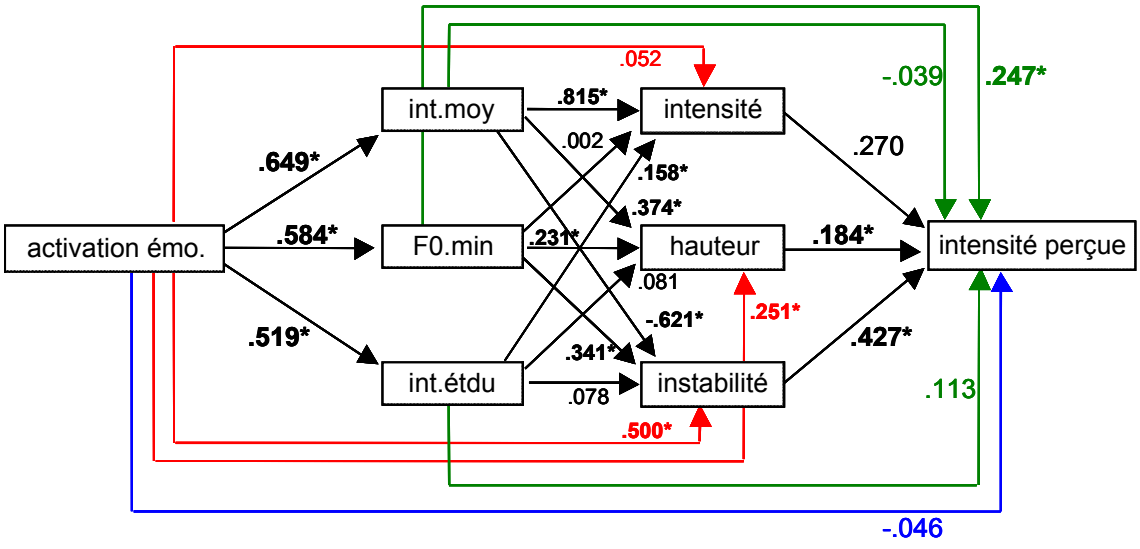
Le tableau 7 présente les résultats d'une régression hiérarchique des trois caractéristiques vocales perçues, puis des trois paramètres acoustiques et, finalement, du niveau d'activation sous-jacent aux émotions exprimées sur l'intensité moyenne perçue. L'introduction des trois paramètres acoustiques dans le modèle (après les caractéristiques vocales perçues) ajoute une contribution significative à l'explication de la variance de l'intensité perçue. En revanche, le niveau d'activation sous-jacent aux émotions exprimées n'améliore pas la prédiction de l'intensité moyenne perçue relativement au modèle qui inclut les trois caractéristiques vocales perçues et les trois paramètres acoustiques.

Tableau 7: Régressions hiérarchiques de 3 caractéristiques vocales perçues, puis de 3 paramètres acoustiques, puis de l'activation exprimée sur l'intensité moyenne perçue.

	Variables acoustiques et dimensions vocales	R	R^2	F change	Sig. F change
intensité perçue	hauteur, instabilité, intensité	0.778	0.605	71.42	.000
	hauteur, instabilité, intensité, int.étdu, F0.min, int.moy	0.804	0.646	5.32	.002
	hauteur, instabilité, intensité, int.étdu, F0.min, int.moy, activation	0.804	0.647	0.30	.588

La figure 19 représente le modèle de path analysis définit pour la relation entre l'activation exprimée et l'intensité perçue. De même que pour les modèles définis pour la communication des émotions, on relève sur ce modèle que les "chemins" (paths) centraux et les "chemins" (paths) périphériques possède des coefficients élevés et significatifs (les coefficients significatifs, $p<.05$, sont indiqués en gras et suivis d'un astérisque). En revanche, contrairement aux modèles définis pour la communication des émotions, le "chemin" direct (représenté par la flèche bleue) correspond à un coefficient non-significatif (statistiquement égal à zéro).

Figure 19: Path analysis pour la relation entre l'activation exprimée et l'intensité moyenne perçue avec 3 caractéristiques vocales et 3 paramètres acoustiques présélectionnés.



Le tableau 8 propose une synthèse des relations représentées dans les figures 15 à 19. La première colonne de ce tableau représente la corrélation entre l'émotion/activation exprimée et l'émotion/intensité perçue. Les colonnes suivantes représentent la somme des effets centraux (deuxième colonne), la somme des effets périphériques basés sur les indices distaux (paramètres acoustiques, dans la troisième colonne), la somme des effets périphériques basés sur les indices proximaux (caractéristiques vocales perçues, dans la quatrième colonne) et l'effet direct entre l'émotion/activation exprimée et l'émotion/intensité perçue (cinquième colonne). La taille

proportionnelle de chaque type d'effet relativement à la corrélation entre l'émotion/activation exprimée et l'émotion/intensité perçue est indiqué en dessous de la valeur correspondant à chaque effet.

Tableau 8: Synthèse de la path analysis, effets centraux et périphériques pour la communication des émotions et pour la relation entre l'activation et l'intensité moyenne perçue

model	Accuracy coefficient		Central effect		Peripheral effects: Distally based		Peripheral effects: Proximally based		Direct effect
	r_{CA}	=	$r_{CD}\rho_{PD}\rho_{AP}$	+	$r_{CD}\rho_{AD}$	+	$\rho_{PC}\rho_{AP}$	+	ρ_{AC}
Colère	0.780	=	0.229	+	(-0.087)	+	0.079	+	0.560
%	100		29		(-11)		10		72
Peur	0.677	=	0.044	+	0.052	+	0.125	+	0.457
%	100		6		8		18		68
Tristesse	0.796	=	0.101	+	(-0.031)	+	0.221	+	0.505
%	100		13		(-4)		28		63
Joie	0.754	=	(-0.009)	+	(-0.004)	+	0.127	+	0.641
%	100		(-1)		(-1)		17		85
Activation	0.577	=	0.172	+	0.177	+	0.274	+	(-0.046)
%	100		30		31		47		(-8)

Des valeurs négatives, relativement proches de zéro, apparaissent dans le tableau 8. Ces valeurs sont représentées entre parenthèses car elles ne sont pas interprétables (en particulier lorsqu'elles sont exprimées sous forme de proportions). Pour l'interprétation, ces valeurs sont considérées comme nulles. Elles présentent l'avantage d'attirer l'attention sur le fait que les valeurs positives proches de zéro doivent de la même manière être considérées comme "non-significatives".

Les résultats présentés dans le tableau 8 peuvent être résumés brièvement en relevant que la proportion de la corrélation médiatisée par les trois caractéristiques vocales perçues et par les trois paramètres acoustiques représente environ un tiers de la corrélation entre l'émotion exprimée et l'émotion perçue pour la colère, la tristesse et la peur; alors qu'une proportion plus faible de la corrélation entre la joie exprimée et la joie perçue (environ un septième) est médiatisée par les trois caractéristiques vocales perçues et par les trois paramètres acoustiques sélectionnés pour cette émotion. En revanche, la totalité de la corrélation entre l'activation exprimée et l'intensité moyenne perçue a pu être médiatisée par les trois caractéristiques vocales perçues et par les trois paramètres acoustiques sélectionnés pour expliquer cette relation. La médiation de la relation entre l'émotion/activation exprimée et l'émotion/intensité perçue ne peut être attribuée assez largement aux effets centraux que dans le cas de la communication de la colère. Une large majorité de la médiation de cette relation passe par les effets périphériques basés sur les indices distaux ou sur les

indices proximaux pour les autres émotions ainsi que pour la relation entre l'activation et l'intensité moyenne perçue.

5.4 Critiques méthodologiques et conceptuelles

Avant de considérer les aspects plus généraux des résultats et d'en tirer des conclusions, quelques points méthodologiques et conceptuels problématiques méritent, à notre avis, d'être exposés et discutés. Deux de ces points concernent des aspects du format des données qui sont liés à la question traitée. Deux autres points ont trait à des problèmes dont l'origine se situe plutôt sur un plan conceptuel.

Les deux points relatifs au format des données incluent d'une part le problème de la colinéarité des paramètres acoustiques et des caractéristiques vocales perçues, et d'autre part le problème de la définition des variables dichotomiques 'émotion/activation exprimée'. Les paramètres acoustiques et les caractéristiques vocales perçues sont par essence colinéaires. Le modèle perceptif lui-même inclut explicitement cet aspect de la redondance des "indicateurs" (cues) présents dans l'environnement (*vicarious functioning*, v. section 1.3). Les résultats présentés ci-dessus indiquent toutefois que la colinéarité de ces variables – utilisées comme prédicteurs dans des régressions multiples – pose un problème relativement à la signification des coefficients de régressions obtenus pour chaque variable. Des coefficients significativement différents de zéro peuvent être très instables au sens où des modifications de faible importance au niveau des valeurs observées peuvent "faire basculer" les valeurs des coefficients. De plus, ces coefficients de régression sont difficiles à interpréter. Ils dénotent la relation entre un prédicteur (paramètre acoustique ou caractéristique vocale perçue) et une variable dépendante (émotion exprimée ou perçue) en présence des autres prédicteurs inclus dans la régression. En conséquence, il nous semble préférable de ne pas mettre un accent trop important sur les coefficients de régressions obtenus. Le modèle basé sur la path analysis s'avèrerait dans ce cas moins indiqué pour l'analyse de la problématique étudiée que le paradigme de la LME qui met un accent moindre sur les coefficients de régression.

Un autre problème lié à la structure des données examinées concerne l'utilisation d'une variable dichotomique pour représenter l'émotion/activation exprimée. A nouveau, il s'agit d'une caractéristique inhérente à la problématique étudiée; chaque type d'émotion (colère, peur, tristesse ou joie) est théoriquement et pratiquement conçu comme étant soit présent, soit absent, dans les expressions utilisées. Or les modèles statistiques appliqués aux données dans la section précédente supposent l'utilisation de variables continues. L'impact exact de l'utilisation d'une échelle dichotomique pour coder l'émotion/activation exprimée est difficile à évaluer; les modèles présentés

pour les quatre émotions et pour l'activation sont toutefois comparables, dans la mesure où ils font tous usage du même codage dichotomique pour représenter l'émotion/activation exprimée.

Deux autres problèmes seront abordés ci-dessous, ils concernent, premièrement, l'utilisation de l'intensité perçue moyenne (dérivée des jugements d'intensité de colère, de joie, de peur et de tristesse) dans le modèle correspondant à l'encodage de l'activation; et deuxièmement, le problème de la sélection empirique des paramètres acoustiques et des caractéristiques vocales perçues pour une partie des modèles présentés.

L'utilisation de la moyenne des jugements d'intensité obtenus pour la colère, la peur, la tristesse et la joie dans les modèles correspondant à l'activation exprimée soulève plusieurs critiques. Premièrement, cette moyenne des jugements d'intensité ne correspond certainement pas aux valeurs qui seraient obtenues si l'intensité émotionnelle était jugée directement (indépendamment du type d'émotion exprimé) par un groupe d'auditeurs. De plus, nous avons amplement souligné la différence conceptuelle qui existe entre l'intensité et l'activation émotionnelle. Dans la section relative aux émotions perçues, il est notamment apparu que l'activation peut ne pas être liée à l'intensité émotionnelle perçue pour certaines catégories d'expressions. Globalement, l'utilisation de la moyenne des jugements d'intensité émotionnelle est donc difficilement justifiable. Cette mesure a toutefois été utilisée ci-dessus, à défaut d'une mesure de l'activation émotionnelle perçue, essentiellement dans le but de démontrer que les caractéristiques vocales, en particulier les paramètres acoustiques, qui figurent dans les modèles présentés permettent d'encoder l'activation exprimée et très probablement de médiatiser également le décodage de l'activation exprimée. Les modèles présentés pour la relation entre l'activation exprimée et l'intensité moyenne perçue doivent cependant être considérés avec prudence dans la mesure ou la relation entre l'activation exprimée et l'activation perçue serait certainement plus importante et la proportion de cette relation médiatisée par les caractéristiques vocales pourrait être moindre.

Un autre aspect de la méthode utilisée ci-dessus soulève également des interrogations relativement à son impact sur les résultats obtenus. L'application des modèles basés sur la "path analysis" a nécessité une réduction du nombre de caractéristiques vocales utilisées dans chaque modèle. Cette réduction a été effectuée en sélectionnant empiriquement trois paramètres acoustiques (relativement à leurs relations observées avec l'émotion exprimée) et trois caractéristiques vocales perçues (pour leurs relations observées avec l'émotion perçue). Cette procédure est évidemment critiquable; relativement à une sélection des caractéristiques vocales qui serait basée sur des hypothèses théoriques, elle permet d'augmenter la proportion de la relation entre l'émotion exprimée et l'émotion perçue qui est médiatisée par les caractéristiques vocales. En revanche, elle ne favorise

pas nécessairement les effets centraux (*central paths*), dans la mesure où, dans la plupart des cas, il n'y a pas de correspondance conceptuelle entre les indices distaux (les paramètres acoustiques) et les indices proximaux (les caractéristiques vocales perçues). Il peut être intéressant de relever que le seul modèle pour lequel une forme de correspondance existe entre les trois paramètres acoustiques et les trois caractéristiques vocales perçues – soit le modèle correspondant à la communication de la colère – est également le seul modèle pour lequel la proportion de la relation médiatisée par les "chemins" (paths) centraux est relativement importante.[17]

Ce dernier point soulève, plus généralement, le problème de la définition et de la sélection des aspects vocaux mesurés. Les paramètres acoustiques et, également, les caractéristiques vocales perçues que nous avons mesurés ont été définis en large partie par des contingences pratiques, plutôt que théoriques. Les résultats obtenus laissent toutefois apparaître que les relations des paramètres acoustiques et des caractéristiques vocales perçues avec l'encodage et le décodage mériteraient d'être mieux spécifiées sur le plan théorique. Ce point sera développé plus en détails ci-dessous dans la section "discussion et conclusions".

5.5 Discussion et conclusions

Deux types de modèles ont été utilisés pour représenter la communication vocale des émotions: le modèle proposé par Juslin, basé sur la LME et le modèle proposé par Scherer, basé sur la path analysis. Les résultats obtenus avec la LME donnent, globalement, une impression plus favorable que les résultats obtenus avec la path analysis. Approximativement 30% de la corrélation entre l'émotion exprimée et l'émotion perçue est expliquée pour la colère, la peur et la tristesse en utilisant la procédure basée sur la path analysis. Alors que les modèles basés sur la LME permettent d'expliquer 58% (en utilisant les paramètres acoustiques) et 71% (en utilisant les caractéristiques vocales perçues) de la relation entre la colère exprimée et la colère perçue; respectivement 41% et 53% de la relation entre la peur exprimée et la peur perçue et 42% et 68% de la relation entre la tristesse exprimée et la tristesse perçue sont expliqués par les paramètres acoustiques et les caractéristiques vocales perçues. Les deux modèles sont toutefois difficilement comparables. Ils proposent non seulement une décomposition différente de la relation entre l'émotion exprimée et l'émotion perçue, mais utilisent également différentes variables: les caractéristiques vocales perçues et les paramètres acoustiques sont utilisés séparément dans les modèles basés sur la LME; alors que

[17] Cette observation est également discutable. La correspondance entre les paramètres acoustiques et les caractéristiques vocales perçues (int.moy/intensité perçue, F0.min/hauteur perçue, v.0-1k/voix perçue comme perçante) est fondée sur une conception plutôt naïve de la relation entre les paramètres acoustiques et les caractéristiques vocales perçues. De plus, la communication de la colère est également mieux modélisée par l'ensemble des paramètres acoustiques et des caractéristiques vocales perçues dans les modèles basés sur la LME.

les modèles basés sur la path analysis ont été obtenus en faisant appel à une sous-sélection de paramètres acoustiques et de caractéristiques vocales perçues (utilisé/es conjointement).

D'autre part, les deux modèles s'accordent plus généralement sur un ensemble de propriétés des résultats – telles que les différences entre les modèles obtenus pour les quatre émotions considérées, les différences observées pour la modélisation de l'encodage et du décodage, la plus grande facilité à rendre compte de l'encodage de l'activation (relativement à l'encodage des émotions), les différences observées entre le pouvoir explicatif des caractéristiques vocales perçues et des paramètres acoustiques. Ces différents aspects sont développés ci-dessous.

Les modèles basés sur la LME, obtenus avec les paramètres acoustiques et avec les caractéristiques vocales perçues, expliquent systématiquement mieux le décodage que l'encodage (R_s>R_e) pour les quatre émotions communiquées. Une interprétation possible de cette observation, serait que les caractéristiques vocales (acoustiques et perçues) sont, *généralement*, utilisées de manière plus cohérente (consistante) au décodage qu'à l'encodage, affectant l'émotion perçue davantage qu'elles ne sont affectées par l'émotion exprimée. Une deuxième interprétation possible, qui n'exclut pas la première, serait que les caractéristiques vocales (acoustiques et perçues) *que nous avons mesurées* sont davantage liées au décodage qu'à l'encodage. Dans cette optique, d'autres caractéristiques pourraient être mesurées qui seraient liées davantage à l'encodage et relativement moins au décodage. La question se pose aussi de l'influence du codage dichotomique de l'émotion exprimée sur ce résultat. Il est en effet possible d'argumenter que la variable dichotomique, dont la distribution est évidemment particulière, pourrait être la source de la relation systématiquement plus faible entre les caractéristiques vocales et l'émotion exprimée (relativement à la relation entre les caractéristiques vocales et l'émotion perçue) qui serait donc un artefact. Le modèle proposé pour l'activation exprimée (tableau 3) offre toutefois une démonstration empirique de l'inexactitude de cette interprétation. La relation entre l'activation exprimée (variable dichotomique) et les caractéristiques vocales est plus forte que la relation entre les caractéristiques vocales et l'intensité moyenne perçue (R_e>R_s). Ce modèle démontre que les caractéristiques vocales peuvent réaliser une très bonne prédiction également à l'encodage (sur la variable dichotomique 'activation exprimée').

Par ailleurs, l'utilisation des caractéristiques vocales perçues relativement à l'utilisation des paramètres acoustiques, dans les modèles LME, produit systématiquement de meilleurs résultats. Cet avantage est particulièrement marqué pour la communication de la joie; pour cette émotion les paramètres acoustiques ne parviennent pas à rendre compte de la communication, alors que les caractéristiques vocales perçues parviennent à expliquer environ un tiers de la relation entre la joie exprimée et la joie perçue. Les modèles basés sur la path analysis corroborent cette observation au

sens où, dans l'ensemble, les effets directs et les effets périphériques basés sur les indices distaux (paramètres acoustiques) sont relativement faibles en comparaison avec les effets périphériques basés sur les indices proximaux (les caractéristiques vocales perçues). Deux réserves doivent toutefois être émises relativement à ces observations. Premièrement, il est possible que d'autres paramètres acoustiques, mieux définis pour représenter la communication vocale des émotions, parviennent à améliorer la modélisation basée sur les paramètres acoustiques. Deuxièmement, dans la mesure où les caractéristiques vocales perçues et les émotions perçues sont des variables qui sont toutes deux évaluées par des procédures de jugements et dans la mesure où les jugements relatifs aux caractéristiques vocales perçues pourraient être influencés par la qualité émotionnelle des expressions, il était prévisible que la relation entre les caractéristiques vocales perçues et les émotions perçues soit plus importante que la relation entre les paramètres acoustiques et les émotions perçues.

En revanche, la relation plus importante entre les caractéristiques perçues et l'émotion exprimée qu'entre les paramètres acoustiques et l'émotion exprimée s'oppose au modèle théorique représenté par figure 2 qui postule que la relation entre l'émotion exprimée et les caractéristiques vocale perçue est médiatisée par les paramètres acoustiques. Les données présentées dans la section 3 de cette thèse laissaient toutefois déjà présager de ce résultat. A la section 3, la discrimination des émotions exprimées réalisées par les caractéristiques vocales perçues, s'est révélée plus performante que la discrimination réalisée par les paramètres acoustiques à la section 2. Dans la discussion présentée à la suite de cette observation (section 3.4), nous avons évoqué la possibilité que la relation plus importante entre les émotions exprimée et les caractéristiques vocales perçues (relativement à la relation entre les émotions exprimées et les paramètres acoustiques) puisse également être expliquée par l'influence de la perception des émotions exprimées sur l'évaluation des caractéristiques vocales perçues (v. section 3.4).

Les différences examinées ci-dessus – entre l'encodage et le décodage, et entre les paramètres acoustiques et les caractéristiques vocales perçues – apparaissent minimes en comparaison avec les différences qui sont apparues entre les modèles correspondant aux quatre émotions et au niveau d'activation. La LME tout comme la path analysis font ressortir en particulier une meilleure modélisation pour la communication de la colère et une moins bonne modélisation pour la communication de la joie. Cette différence importante – encore renforcée par la possibilité de modéliser la communication de l'activation-intensité d'une manière presque parfaite – démontre que des caractéristiques vocales qui peuvent rendre compte de la communication de certaines émotions (ou de l'activation) ne parviennent pas nécessairement à rendre compte d'autres émotions. Cette observation met donc l'accent sur la nécessité d'identifier des caractéristiques vocales plus

spécifiques aux différents types d'émotions exprimées. Plus spécifiquement, la modélisation peu performante obtenue pour la communication de la joie suggère que les caractéristiques vocales que nous avons mesurées échoueraient à rendre compte de la distinction entre une valence positive et une valence négative communiquée par la voix. Curieusement, ce résultat et la proposition qui en découle ne sont que très rarement évoqués dans les études de la communication vocale des émotions. Dans une revue de la littérature incluant également des résultats obtenus pour la voix chantée, Sundberg (1987) propose toutefois une hypothèse similaire: "it is perhaps not only by duration of syllables and pauses, by voice amplitude, and by rise and decay time of the tones that we express happiness in singing. Still these parameters are involved in the patterns typical of sorrow, fear, anger, and a neutral state" (Sundberg, 1987, p. 153).

Plus généralement, pour les quatre émotions communiquées, les modèles basés sur la path analysis mettent en évidence le fait que l'émotion perçue partage une variance plus importante avec l'émotion exprimée, qu'avec les caractéristiques vocales (acoustiques et perçues). En conséquence, l'émotion exprimée est, pour les quatre émotions, le meilleur prédicteur de l'émotion perçue. Cet aspect des données est problématique pour les modèles basés sur la path analysis qui supposent que les variables proximales (les caractéristiques vocales perçues) sont les variables les plus directement liées à l'émotion perçue. L'émotion exprimée explique de plus une part relativement importante de la variance de l'émotion perçue qui n'est pas expliquée par les caractéristiques vocales perçues et par les paramètres acoustiques. Cet aspect est représenté dans le tableau 5 qui met en évidence le fait que, dans des régressions hiérarchiques, l'émotion exprimée augmente significativement la variance prédite de l'émotion perçue lorsqu'elle est ajoutée à un modèle défini à l'aide de trois caractéristiques vocales perçues et de trois paramètres acoustiques. Le tableau 7 indique en revanche que le niveau d'activation lié aux émotions exprimées ne contribue pas significativement à expliquer la variabilité de l'intensité moyenne perçue lorsqu'il est ajouté à un modèle défini à l'aide de trois caractéristiques vocales perçues et de trois paramètres acoustiques.

L'ensemble des résultats implique en fait que les paramètres acoustiques et les caractéristiques vocales perçues qui sont utilisées dans les différents modèles présentés ci-dessus ne parviennent à représenter qu'une part relativement faible de la communication émotionnelle, alors qu'ils représentent probablement efficacement la communication de l'activation. Le problème central semble donc bien se situer au niveau du choix et de la définition des paramètres acoustiques mesurés et des caractéristiques vocales perçues évaluées. Dans cette étude, les 44 paramètres acoustiques initialement extraits des signaux correspondent à des mesures qui sont habituellement effectuées dans le domaine de la recherche sur la communication vocale des émotions. Ils sont limités à des mesures qui peuvent être relativement facilement (et de manière semi-automatique)

obtenues à l'aide des logiciels d'analyse acoustiques actuels. L'utilisation d'autres paramètres acoustiques – tels que des analyses de formants ou l'évaluation du *jitter* – ont été envisagés, mais ils sont en pratique difficiles à extraire pour les expressions que nous avons analysées[18]. Il existe donc une tension entre la nécessité évidente de mieux spécifier les caractéristiques acoustiques qui interviennent probablement dans la communication vocale des émotions – en se fondant notamment sur des considérations théoriques – et les mesures qu'il est, à ce jour, possible d'obtenir. La même observation s'applique également pour les caractéristiques acoustiques perçues. On souhaiterait idéalement pouvoir obtenir des évaluations relatives à d'autres caractéristiques que celles qui ont été évaluées dans cette étude. Malheureusement, les caractéristiques qui pourraient être intéressante pour la communication de certaines émotions – par exemple la qualité soufflée (*breathiness*) qui pourrait être importante pour la communication de la peur/anxiété, ou encore la présence d'une traction latérale des lèvres (*lip-spreading*) qui pourrait être importante pour la communication de la joie/amusement – ne peuvent probablement pas être évaluées fidèlement par des auditeurs qui ne sont pas préalablement entraînés.

Ces considérations ne nous empêchent toutefois pas de formuler quelques suggestions relativement aux aspects vocaux qui devraient être mesurés afin d'obtenir des indicateurs (cues) plus adaptés pour représenter la communication vocale des émotions. En premier lieu, il semblerait surtout approprié de mesurer des caractéristiques acoustiques dont la relation avec la production vocale et/ou la perception serait mieux spécifiée. A l'heure actuelle les mesures disponibles dans les logiciels d'analyses acoustiques permettent d'extraire des paramètres qui décrivent les signaux mais qui ne sont pas orientés spécifiquement vers la production vocale ou vers la perception; un effort de recherche et de développement important devrait donc être entrepris dans cette direction. Un aspect qui nous semble intéressant concerne notamment la description de l'intonation et de son rôle dans la communication vocale des émotions. La F0 moyenne et l'intensité moyenne (ainsi que leurs étendues ou leurs écarts-types) représentent probablement des indicateurs trop rudimentaires et ne résument sans doute pas adéquatement les fluctuations des courbes de F0 et d'intensité. Le dernier chapitre de cette thèse sera notamment consacré à l'analyse détaillée des fluctuations de la F0.

Enfin, l'aspect des interactions possibles entre différentes caractéristiques vocales (acoustiques et/ou perçues) mériterait d'être examiné avec plus d'attention. Les modèles linéaires présentés ci-dessus ne laissent, en effet, pas de place aux configurations qui pourraient éventuellement jouer un rôle important dans la communication vocale des émotions. Evaluer la présence ou l'effet de

---

[18] Les expressions émotionnelles simulées qui ont été utilisée dans cette étude ne comportent pas de voyelles soutenues. De plus l'expression de l'émotion "dégrade" la qualité de l'articulation et de la parole, plus généralement. Il est en conséquence assez délicat de réaliser des descriptions phonétiques poussées pour ce type d'expressions.

configurations spécifiques entre des dimensions prédéfinies requiert le développement et la formulation d'hypothèses quant aux associations/combinaisons entre les caractéristiques vocales qui découleraient de l'expression d'une émotion et qui pourraient être éventuellement utilisées pour identifier cette émotion. De telles hypothèses sont quasi inexistantes dans ce domaine de recherche et mériteraient d'être développées dans des études ultérieures.

6 Analyse de la contribution de l'intonation à la communication des émotions

6.1 Introduction

Dans les chapitres précédents, un ensemble de paramètres acoustiques ont été mesurés et ont été mis en relation avec les émotions exprimées et avec les émotions perçues dans des expressions vocales. Ces paramètres ont été choisis en référence aux paramètres utilisés dans d'autres études qui se sont intéressées aux caractéristiques acoustiques des expressions vocales émotionnelles (en particulier Banse & Scherer, 1996). Dans la plupart des études qui ont été réalisées à ce jour, on trouve des mesures ayant trait à la durée, à la fréquence fondamentale et à l'intensité des expressions. Des mesures qui résument les contours de F0 et/ou les contours d'intensité au niveau des énoncés – telles que la moyenne et l'écart-type, plus rarement le minimum, le maximum et l'étendue – sont en général rapportées pour chaque expression/énoncé étudié, puis pour chaque émotion exprimée. D'autres paramètres sont également utilisés mais leur usage est moins répandu. Dans quelques études, des analyses spectrales ont notamment été effectuées. Banse & Scherer (1996) ont par exemple mesuré la proportion d'énergie dans différentes bandes spectrales, alors que dans d'autres travaux l'effet de l'émotion sur les formants a été évalué (par exemple Tolkmitt, Helfrich, Standke, & Scherer, 1982).

Dans la littérature aussi bien que dans l'étude présentée ci-dessus (v. sections 2 et 5), l'analyse acoustique se limite donc le plus souvent à l'extraction de paramètres segmentaux qui sont en général agrégés pour la totalité d'un énoncé ou parfois pour des sections de parole plus importantes. Les résultats rapportés dans les chapitres précédents indiquent que ces paramètres acoustiques ne parviennent à rendre compte que d'une partie des différences entre les émotions exprimées et n'expliquent également qu'une partie des attributions émotionnelles effectuées par des auditeurs.

Différents auteurs ont par ailleurs affirmé de manière répétée que parmi les aspects prosodiques de la parole, l'intonation joue un rôle prépondérant dans la communication vocale des émotions (v. section 1.4 de cette thèse). Or l'intonation correspond à des aspects prosodiques (tels que l'accentuation et le rythme de parole) qui sont liés au décours temporel des expressions et qui sont en grande partie supprimés par l'utilisation de moyennes et d'indices de variabilité calculés au niveau des énoncés.

Les modèles qui ont été proposés pour l'étude de l'intonation, ainsi qu'une revue de quelques propositions et de quelques résultats empiriques qui défendent l'importance du rôle de l'intonation

dans la communication vocale émotionnelle ont été présentés dans l'introduction théorique de cette thèse (section 1.4). L'approche que nous avons appliquée est présentée dans la section qui suit.

6.1.1 Approche choisie, aspects de l'intonation examinés

Conformément à la tendance dominante dans les travaux issus de la recherche en psychologie et en linguistique, nous avons choisi d'examiner uniquement les contours de fréquence fondamentale (à l'exclusion des contours d'intensité) des expressions émotionnelles qui ont été décrites dans les sections précédentes.

Dans sa thèse de 1998, Mozziconacci a présenté des résultats concernant la description de la hauteur perçue d'un ensemble d'expressions émotionnelles produites par des acteurs en utilisant le système de transcription en contours de l'IPO, basé sur le modèle de l'intonation du hollandais développé par t'Hart et al. (1990). Dans la thèse de Mozziconacci, une stylisation des contours de F0 – sous forme d'un codage de la hauteur (en semi-tons et en Hz) pour 6 points d'encrage dans les énoncés – a également été réalisée pour un ensemble d'expressions émotionnelles produites par 3 acteurs. Les points relevés sur les contours de F0 ne sont toutefois utilisés que pour définir différentes estimations de l'étendue de la F0 (différence entre un maxima et différents minima locaux). À notre connaissance, aucune étude n'a jusqu'ici comparé systématiquement d'autres propriétés des contours de F0 pour différentes expressions émotionnelles produites par des acteurs.

Les systèmes linguistiques qui sont basés sur la transcription de la hauteur perçue – tels que, par exemple le système ToBI (v. section 1.4) ou le système de l'IPO – n'ont pas été retenus pour l'étude présentée ci-dessous pour les raisons suivantes: (a) Ces systèmes proposent des catégories (un nombre limité de tons ou de courbes) pour la transcription de la hauteur qui ont été développées pour la description de l'intonation linguistique. Etant donné la spécificité de la parole émotionnelle, on peut, raisonnablement, mettre en doute la pertinence de ces catégories pour la transcription de l'intonation émotionnelle. (b) La hauteur perçue qui est transcrite est influencée non seulement par les variations de la F0, mais également par les variations de la durée, de l'intensité et de la distribution spectrale de l'énergie des expressions. Les transcriptions de hauteur perçues n'apportent pas d'information quant aux aspects acoustiques qui sont à l'origine de l'impression auditive transcrite. (c) Des travaux (e.g. Syrdal & McGory, 2000; Wightman, 2002) démontrent que ces transcriptions sont relativement peu fiables; les transcriptions varient selon les individus qui les réalisent. En conséquence, on peut formuler l'hypothèse qu'elles sont également susceptibles d'être influencées par la perception de la qualité émotionnelle de la voix. (d) Afin de comparer un grand nombre d'énoncés produits indépendamment et qui expriment différentes émotions, un point de référence extérieur aux énoncés doit être fixé. Or les transcriptions de la hauteur perçues sont en

principe réalisées sans point de référence extérieur, les variations perceptibles de la hauteur sont transcrites relativement au niveau général de la séquence considérée. (e) Plus généralement, ces systèmes proposent des descriptions en catégories qui ne permettent pas de quantifier les différences observées entre différents contours de hauteurs transcrits.

Cette série de critiques adressées aux systèmes de transcription de la hauteur perçue permet de mettre en évidence les points que nous avons choisi de privilégier. Premièrement, dans l'étude présenté ci-dessous, le codage de l'intonation est orienté vers la production. Au long des chapitres précédents, la production (encodage) et la perception (décodage) des caractéristiques vocales ont été soigneusement distinguées. La perspective adoptée ici soutient qu'il est important d'examiner d'abord la manière dont les émotions exprimées affectent les caractéristiques vocales examinée (dans le cas présent le contour de F0) et, dans un deuxième temps, la manière dont ces caractéristiques affectent la perception et la reconnaissance des émotions exprimées. Deuxièmement, le système de codage est quantitatif avec un point de référence externe qui permettra de comparer un ensemble de propriétés du contour de F0 pour différents énoncés produits séparément. Troisièmement, le système de codage se base sur une série de critères objectifs et minimise ainsi la part de l'interprétation due au codeur.

Dans cette perspective, une approche non contrainte par un modèle linguistique a été définie sous la forme d'une stylisation systématique des contours de F0. Cette stylisation systématique a été effectuée en relevant des points cruciaux du contour de F0 (maxima et minima locaux) pour chaque énoncé. La définition de cette procédure a été orientée par les conseils de D. R. Ladd (communication personnelle) et par une étude de Patterson & Ladd (1999). Les détails concernant les critères utilisés pour effectuer la stylisation des contours sont présentés dans la section méthode ci-dessous. Les aspects du contour de F0 qui varient avec les émotions exprimées ont été examinés et ont permis de générer un petit nombre d'hypothèses concernant les aspects de la hauteur perçue qui pourraient affecter les attributions émotionnelles chez des auditeurs. L'effet des variations de F0 sur les attributions émotionnelles a été testé dans un deuxième temps en évaluant l'impression émotionnelle produite par des expressions de synthèse dont les contours de F0 ont été systématiquement modifiés. Les manipulations de la parole (synthèse et resynthèse) ont été effectuées pas Michel Morel à l'Université de Caen dans le cadre d'une collaboration initiée par le Projet Plurifacultaire Prosodie (financé par l'Université de Genève). Des études de jugements, réalisées à l'Université de Genève ont permis d'évaluer dans quelle mesure les manipulations de l'intonation affectent les attributions émotionnelles réalisées par des auditeurs. Les détails concernant les manipulations de l'intonation et les procédures de jugement utilisées, ainsi que les résultats obtenus sont présentés dans la section 6.3 ci-dessous.

6.2 Stylisation des contours de F0

Les caractéristiques du contour de la F0 qui pourraient être liées à l'expression de différentes émotions sont à ce jour très mal connues. Afin de décrire systématiquement et de comparer les contours de F0 pour un ensemble d'expressions émotionnelle, une stylisation manuelle des contours a été réalisée en identifiant certains points jugés importants a priori. Plus spécifiquement, des excursions de F0 ("accents") attendues sur des segments phonétiques prédéfinis ont été relevées. Les critères utilisés pour ce codage (stylisation) de la F0 sont décrits ci-dessous.

6.2.1 Méthode: procédure utilisée pour la stylisation

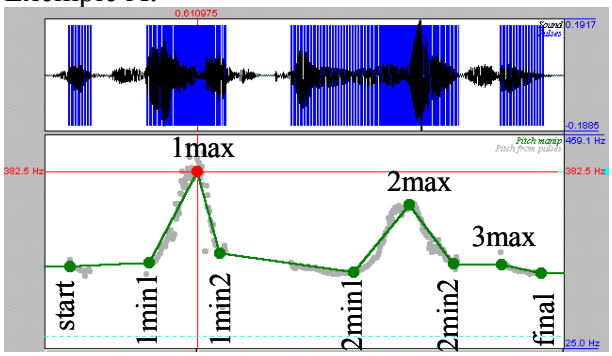
La stylisation manuelle du contour de la fréquence fondamentale a été effectuée pour 144 enregistrements. Ces enregistrements ont été extraits d'une base de données constituée et décrite en détail par Banse & Scherer (1996). Des enregistrements produits par 9 acteurs ont été sélectionnés. Tous les acteurs prononcent 2 séquences de 7 syllabes sans signification (1. "hät san dig prong nju ven tsi", 2. "fi götte laich jean kill gos terr") et expriment 8 types d'émotions : colère chaude ('rage') et colère froide ('irrit'), anxiété ('anx') et peur panique ('paniq'), tristesse ('trist') et désespoir ('desp'), joie calme ('joie') et joie intense ('elat'). La fréquence fondamentale a été extraite par auto-corrélation pour chacune des 144 expressions émotionnelles à l'aide du logiciel PRAAT (Boersma & Weenink, 1996). Les contours extraits ont été contrôlés manuellement. Les erreurs de calcul correspondant à la détection d'une période de F0 sur des parties non voisées ont été corrigées.

Dix points de chaque contour de F0 devaient en principe été relevés pour chaque enregistrement. Le premier point ('start') correspond à la hauteur initiale de la première partie voisée de chaque séquence, c'est-à-dire à la première valeur de F0 détectée pour la syllabe "hät" dans la première séquence de syllabes et à la première valeur de F0 détectée pour la syllabe "fi" dans la deuxième séquence de syllabes. L'absence de détection d'une période de F0 et les erreurs de détection sur ces syllabes ont été enregistrées comme données manquantes. Les deuxième ('1min1'), troisième ('1max') et quatrième points ('1min2') correspondent respectivement aux minimum, maximum, minimum de l'excursion de F0 pour le premier "accent" de chaque séquence. Ces minima et maxima locaux ont été relevés pour les syllabes "san dig" dans la première séquence de syllabes. Pour la deuxième séquence, ces valeurs sont relevées sur les syllabes "götte laich". Les points cinq, ('2min1') six ('2max') et sept ('2min2') correspondent respectivement aux minimum, maximum, minimum de l'excursion de F0 pour le deuxième "accent" de chaque séquence. Ils ont été relevés pour les syllabes "prong nju ven" et "jean kill gos". Pour chaque excursion, le premier minimum correspond au point où la pente de la fréquence fondamentale devient positive. Au cas où une forte augmentation de la pente est précédée d'une section plate ou avec une pente positive très faible,

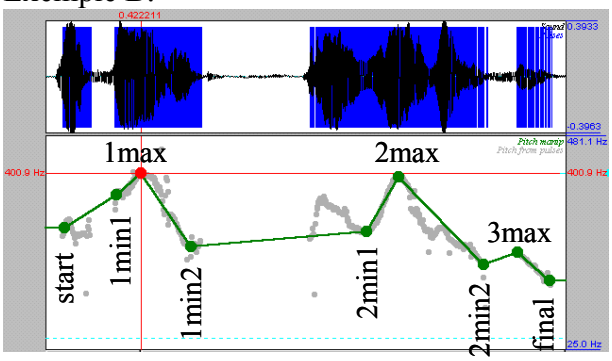
cette section est ignorée. Le maximum correspond au point où la pente de la fréquence fondamentale devient négative. Les fluctuations légères de la pente sont ignorées. Le deuxième minimum correspond au point où la pente de la fréquence fondamentale n'est plus négative. A nouveau, les fluctuations légères – par exemple une légère montée locale suivie par un prolongement de la descente de F0 – sont ignorées. Lorsqu'une pente descendante forte est suivie par une section plate ou très légèrement descendante, cette section est ignorée. Les points huit ('3min'), neuf ('3max') et dix ('final') correspondent à "l'accent final" de chaque séquence ; les minimum, maximum, minimum locaux sont relevés pour les syllabes "tsi" et "ter". Sur ces dernières syllabes, il est relativement rare d'observer une montée suivie d'une descente. Le plus souvent, on observe uniquement une descente finale de la F0. Dans ce cas le point huit (premier minimum de "l'accent final") et considéré comme donnée manquante. Lorsque la montée est absente sur les groupes syllabiques "san dig", "prong nju ven", "gött laich", "jean kill gos", les points 2 et 5 (premiers minima des deux "accents") sont également notés comme données manquantes. Lorsque au contraire on observe uniquement une montée de la F0 sur les différents groupes syllabiques considérés, les points 4, 7 et 10 sont considérés comme données manquantes. La figure 1 présente des illustrations de ce codage pour trois expressions émotionnelles. Les contours de F0 originaux sont représentés en gris, les contours stylisés sont surimposés en vert/noir. Le point 8 (minimum précédant la descente finale sur la dernière syllabe) est absent dans les exemples A et B.

Figure 1: Illustration du codage du contour de F0 pour trois expressions émotionnelles

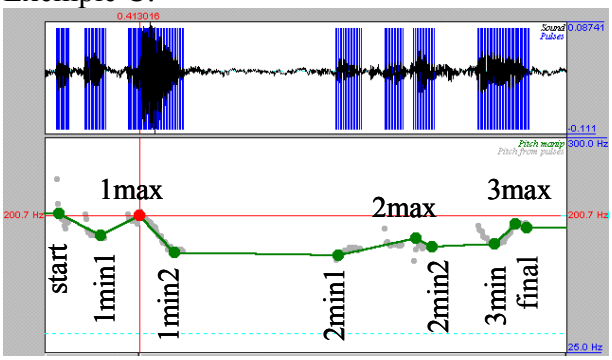
Exemple A:



Exemple B:



Exemple C:



L'**exemple A** (une expression de joie calme, avec la 1ère séquence de syllabes) illustre un cas qui ne présente pas de difficultés particulières.

L'**exemple B** (une expression de colère chaude avec la 1ère séquence de syllabes) illustre un cas où une excursion locale sur la 2ème partie de l'expression a été ignorée.

L'**exemple C** (une expression de joie calme avec la 2ème séquence de syllabes) illustre un cas douteux où deux excursions de faible amplitude ont été codées. Deux montées sur la 2ème et la 3ème partie de l'expression correspondraient à un codage plus économique.

Ce codage de la fréquence fondamentale permet de définir le type d'excursion (une montée suivie d'une descente, une descente ou une montée uniquement, l'absence d'excursion) réalisé sur un segment de parole prédéfini ; dans le cas présent, il s'agit en partie de segments ou un accent est attendu. Il permet également de décrire la durée (en secondes) et l'amplitude (en Hz) de chaque excursion de F0, ainsi que la hauteur absolue (en Hz) des "accents" pour différentes expressions émotionnelles. Ces valeurs peuvent être directement comparées pour un ensemble d'expressions émotionnelles produites par un même locuteur, mais doivent être ajustées relativement aux valeurs caractéristiques de la voix de chaque locuteur, lorsque des expressions produites par plusieurs locuteurs sont comparées.

6.2.2 Résultats

6.2.2.1 Fréquence des excursions observées en fonction de l'émotion exprimée, du locuteur et de la phrase

Les résultats de la stylisation de la F0 sont présentés ci-dessous. Dans un premier temps les types d'excursions observées sont décrites pour chacun des segments considérés en fonction de l'émotion exprimée (8 types d'émotion), en fonction du locuteur (9 acteurs) et en fonction de la "phrase" (1ère ou 2ème séquence de syllabe). Pour le premier point observé (première syllabe de chaque "phrase"), le point peut être présent ou absent. Pour les trois autres parties sur lesquelles des excursions sont attendues ('acc1', 'acc2' et 'final'), on peut observer un "accent" (c'est-à-dire la présence de trois valeurs: 'min1'-'max'-'min2/final' qui correspondent à une montée suivie d'une descente), une montée ('min1'-'max'), une descente ('max'-'min2/final') ou l'absence de l'excursion qui correspond soit à l'absence de F0 détectée sur ce segment, soit à une section plate du contour de F0. Ces résultats sont présentés dans trois tableaux successifs : Le tableau 1 présente les résultats par émotion exprimée, le tableau 2 présente les résultats en fonction du locuteur et le tableau 3 présente les résultats selon la "phrase" prononcée.

En ce qui concerne l'observation des différentes excursions de F0 en fonction de l'émotion exprimée (tableau 1), le chi^2 de Pearson indique que les types d'excursions pour les différents segments sont indépendants du type d'émotion exprimée: Pour la première syllabe (start), $chi^2(7) = 12.34$, $p = .090$; pour le premier "accent", $chi^2(21) = 22.71$, $p = .360$; pour le deuxième "accent", $chi^2(21) = 21.56$, $p = .425$; pour la dernière syllabe (final), $chi^2(21) = 30.50$, $p = .082$. Le nombre d'observations est en principe insuffisant pour l'application de cette méthode statistique, les valeurs rapportées ci-dessus, ainsi que les autres estimations du chi^2 de Pearson présentées ci-dessous, sont données à titre indicatif uniquement.

Tableau 1: Types d'excursion de la F0 par segment et par émotion exprimée. Le nombre d'excursions de chaque type est rapporté pour chaque émotion exprimée.

segment	excursion F0	anx	joie	rage	irrit	paniq	trist	elat	desp	moyenne
start	pt observé	14	16	18	18	16	14	18	16	16,25
	absent	4	2	0	0	2	4	0	2	1,75
acc1	accent	10	11	12	8	9	9	12	11	10,25
	montée	5	4	4	8	8	5	5	1	5,00
	descente	1	1	0	2	1	3	0	2	1,25
	absent	2	2	2	0	0	1	1	4	1,50
acc2	accent	16	16	15	16	16	11	15	13	14,75
	montée	0	0	1	0	1	1	1	1	0,63
	descente	1	0	0	1	0	5	1	2	1,25
	absent	1	2	2	1	1	1	1	2	1,38
final	accent	4	5	4	5	6	1	9	6	5,00
	montée	0	2	1	0	1	3	1	0	1,00
	descente	8	9	11	11	9	6	6	10	8,75
	absent	6	2	2	2	2	8	2	2	3,25

En ce qui concerne l'observation des différentes excursions de F0 en fonction du locuteur (tableau 2), le chi^2 de Pearson indique que les types d'excursion sur le premier "accent" et sur le deuxième "accent" sont dépendants du locuteur: Pour le premier "accent", $chi^2(24) = 54.77$, $p < .001$; pour le deuxième "accent", $chi^2(24) = 39.21$, $p = .026$. Les types d'excursions réalisés sur la première syllabe (start) et sur la dernière syllabe (acc fin) sont indépendants du locuteur: pour la première syllabe, $chi^2(8) = 8.70$, $p = .368$; pour la dernière syllabe, $chi^2(24) = 31.30$, $p = .145$.

Tableau 2: Types d'excursion de la F0 par segment et par locuteur. Le nombre d'excursions de chaque type est rapporté pour chaque locuteur.

segment	excursion F0	loc2	loc3	loc4	loc7	loc8	loc9	loc10	loc11	loc12	moy.
start	pt observé	15	14	13	16	13	15	16	15	13	14,44
	absent	1	2	3	0	3	1	0	1	3	1,56
acc1	accent	5	7	9	7	15	13	6	12	8	9,11
	montée	7	4	5	8	1	1	7	4	3	4,44
	descente	3	1	1	0	0	2	3	0	0	1,11
	absent	1	4	1	1	0	0	0	0	5	1,33
acc2	accent	12	11	13	11	15	16	13	15	12	13,11
	montée	1	2	0	0	0	0	2	0	0	0,56
	descente	2	0	1	4	1	0	1	1	0	1,11
	absent	1	3	2	1	0	0	0	0	4	1,22
final	accent	2	5	3	6	7	4	6	3	4	4,44
	montée	0	1	1	0	1	3	1	0	1	0,89
	descente	9	6	8	9	7	7	8	12	4	7,78
	absent	5	4	4	1	1	2	1	1	7	2,89

Finalement, en ce qui concerne les différents types d'excursions observés selon la "phrase" prononcée (tableau 3), le chi^2 de Pearson indique que les types d'excursion sur le premier "accent" et sur la syllabe finale sont dépendants de la phrase prononcée: Pour le premier "accent",

$chi^2(3) = 10.75$, $p = .013$; pour la dernière syllabe, $chi^2(3) = 45.95$, $p < .001$. Les types d'excursions réalisés sur la première syllabe (start) et sur le deuxième "accent" sont indépendants de la "phrase" prononcée: Pour le deuxième "accent", $chi^2(3) = 7.10$, $p = .069$; pour la première syllabe, $chi^2(1) = 0.32$, $p = .574$.

Tableau 3: Types d'excursion de la F0 par segment et par "phrase". Le nombre d'excursions de chaque type est rapporté pour chaque "phrase".

segment	excursion F0	phrase 1	phrase 2	moyenne
start	pt observé	64	66	65,00
	absent	8	6	7,00
acc1	accent	50	32	41,00
	montée	12	28	20,00
	descente	4	6	5,00
	absent	6	6	6,00
acc2	accent	64	54	59,00
	montée	1	4	2,50
	descente	5	5	5,00
	absent	2	9	5,50
final	accent	3	37	20,00
	montée	2	6	4,00
	descente	49	21	35,00
	absent	18	8	13,00

Le faible nombre d'observations ne permet pas de tirer des conclusions définitives. Toutefois, les différences concernant le type d'excursion réalisé sur les différents segments des expressions semblent apparaître surtout pour les locuteurs et pour les "phrases". Les émotions exprimées, en revanche, ne semblent affecter le type d'excursion codé pour aucun des quatre segments considérés.

6.2.2.2 Effet de l'émotion exprimée sur la hauteur des différents points relevés

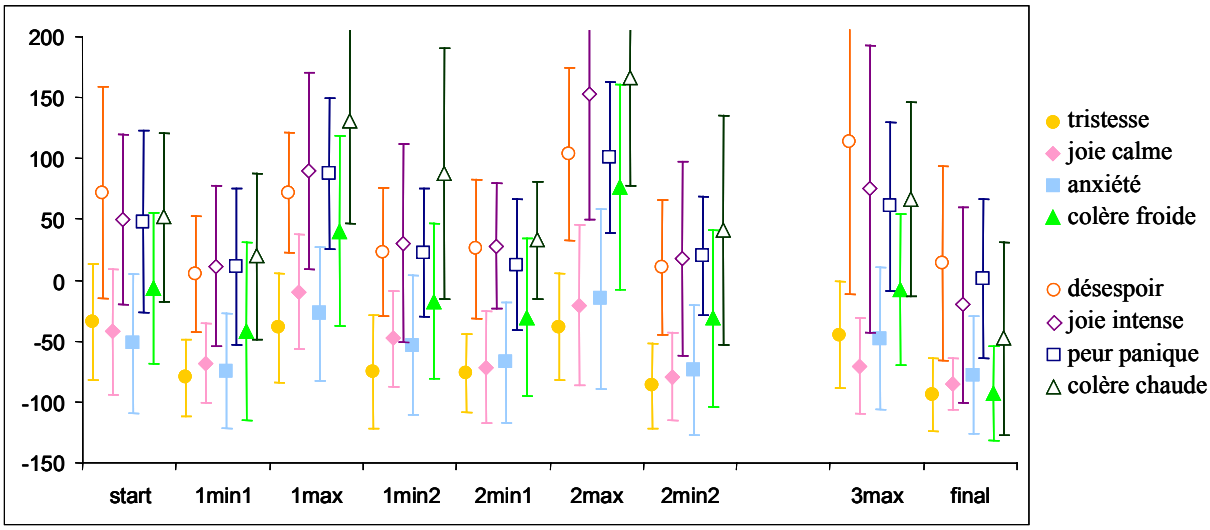
Etant donné la grande quantité de points de F0 non codés (en l'absence de détection ou en l'absence d'excursion de la F0), les effets de l'émotion exprimée sur les différents points relevés ne peuvent être évalués simultanément aux effets de la "phrase" et du locuteur.[19] Bien que les points mesurés varient sans conteste en fonction des "phrases" et des locuteurs, c'est l'effet de l'émotion exprimée qui représente le centre d'intérêt de cette étude. Cet effet sera donc examiné ci-dessous dans le contexte de la variabilité due aux 2 "phrases" et aux 9 locuteurs. Le niveau de F0 propre à chaque locuteur a toutefois été contrôlé de manière à limiter, autant que possible, la confusion entre un effet de l'émotion et un effet du locuteur qui pourrait survenir dans un contexte où les observations ne

[19] Sur les 144 expressions étudiées, il n'y a, au maximum, qu'une observation pour chaque cas de l'interaction émotion(8) * phrase(2) * locuteur(9) et deux observations pour l'interaction émotion * locuteur. Lorsque des observations manquent, ces interactions ne peuvent être évaluées.

sont pas également réparties entre les différents locuteurs et les différentes émotions. Ce contrôle du niveau de F0 a été réalisé en définissant une F0 moyenne pour chaque locuteur sur la base de 112 expressions produites par chaque locuteur (incluant les 16 expressions analysées). Cette valeur moyenne spécifique à chaque locuteur a été soustraite des valeurs de F0 relevées lors de la stylisation des contours.

Le graphique 1 représente les moyennes et les écarts-types des valeurs de F0 (en Hz) pour chaque point codé en fonction du type d'émotion exprimée, après soustraction de la F0 moyenne (en Hz) définie pour chaque locuteur. Ces valeurs moyennes ne sont pas représentées pour le point '3min' qui n'a été observé que pour 48 des 144 expressions et seulement cinq fois pour la première "phrase" (v. tableaux 1 à 3). La F0 moyenne par émotion exprimée pour chacun des 10 points des contours, ainsi que les écarts-types et le nombre d'observations, sont également représentés dans deux tableaux en annexe (E.1 et E.2). Le premier tableau (annexe E.1) représente les valeurs avant la soustraction de la F0 moyenne définie pour chaque locuteur; le deuxième tableau (annexe E.2) représente les valeurs après la soustraction de la F0 moyenne définie pour chaque locuteur.

Graphique 1: Moyennes et écarts-types par type d'émotion exprimée pour les points de F0 codés



Ce graphique met en évidence, d'une part, la très forte variabilité pour les expressions correspondant à un même type d'émotion. L'écart-type moyen, toutes émotions et points confondus, est égal à 62 Hz (avec un maximum de 125 Hz pour le point '3max' et les expressions de 'désespoir', et un minimum de 21 Hz pour le point 'final' et les expressions de 'joie calme'). D'autre part, ce graphique indique que la plupart des différences entre les émotions exprimées sont imputables au niveau d'activation sous-jacent aux réactions émotionnelles. Les moyennes pour les émotions avec une faible activation – joie calme, tristesse, anxiété, colère froide – sont plus faibles que les moyennes des émotions avec une forte activation – joie intense, désespoir, peur panique, colère chaude – pour l'ensemble des points codés.

Une analyse de variance a été effectuée sur les valeurs de F0 codées (après soustraction de la moyenne définie pour chaque locuteur) en utilisant deux facteurs: l'émotion exprimée (8 niveaux) et le point relevé (9 niveaux – les valeurs relevées pour le point '3min' ont été exclues de cette analyse). Cette analyse ne tient pas compte de la dépendance qui existe entre les points de F0 successivement relevés sur les mêmes expressions, ni de la dépendance des mesures liée au fait que les mêmes locuteurs ont exprimé les différents types d'émotion. Les résultats de cette ANOVA indiquent qu'il existe un effet principal de l'émotion exprimée sur les valeurs de F0 mesurées ($F(7, 1016) = 86.45$, $p < .001$, $\eta^2 = .37$). Le test post-hoc HSD de Tukey indique que cette différence est due au fait que les moyennes générales des points relevés pour la tristesse, la joie et l'anxiété sont significativement plus faibles que la moyenne pour la colère froide; la moyenne pour la colère froide étant elle-même significativement plus faible que les moyennes pour la peur panique, la joie intense, le désespoir et la colère chaude. L'analyse indique également qu'il existe un effet principal du point mesuré ($F(8, 1016) = 37.86$, $p < .001$, $\eta^2 = .23$), cet effet reflète directement la méthode utilisée pour le codage de la F0. Les moyennes pour les minima relevés sont plus faibles que les moyennes pour les maxima ; la moyenne pour le point initial ('start') se situe entre les moyennes obtenues pour les maxima et les minima. Les différences significatives selon le test post-hoc HSD de Tukey pour l'émotion exprimée et le point relevé sont représentées dans le tableau 4.

Tableau 4: Test post-hoc (Tukey HSD), différences significatives ($p<.05$) entre les moyennes pour les 8 émotions exprimées et pour les 9 points relevés dans les contours.

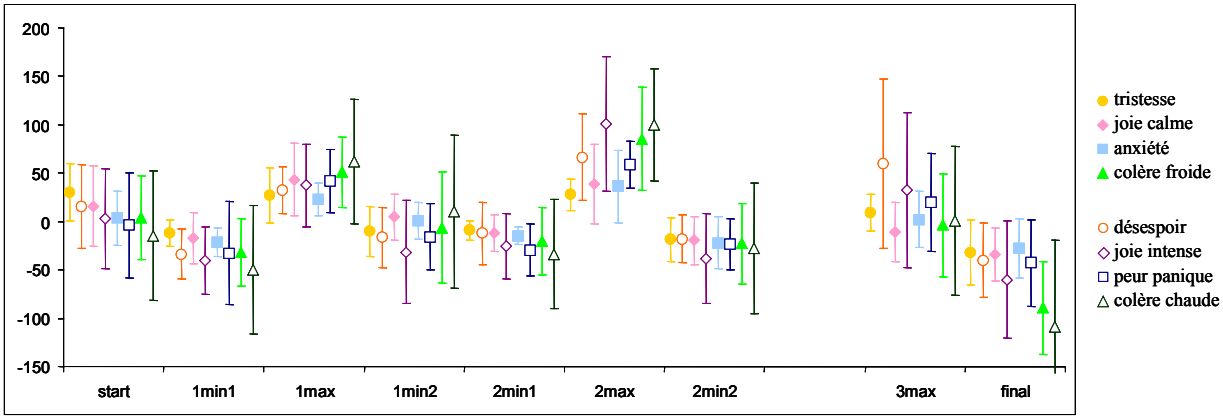
EMO (moy)					POINT (moy)			
trist (-60 Hz) joie (-54 Hz) anx (-53 Hz)	<	irrit (-10Hz)	<	paniq (43 Hz) exalt (51 Hz) desp (51 Hz) rage (62 Hz)	final (-45 Hz)	<	2min1 (-16Hz) 1min2 (-3 Hz) start (14 Hz)	3max (25 Hz) 1max (43 Hz) 2max (66 Hz)
					1min1 (-26 Hz) 2min2 (-23 Hz) 2min1 (16Hz)	<	start (14 Hz)	3max (25 Hz) 1max (43 Hz) 2max (66 Hz)
					1min2 (-3 Hz)	<	3max (25 Hz) 1max (43 Hz) 2max (66 Hz)	
					start (14 Hz)	<	1max (43 Hz) 2max (66 Hz)	
					3max (25 Hz)	<	2max (66 Hz)	

L'interaction entre l'émotion exprimée et le point relevé est également significative, mais la taille de cet effet est beaucoup plus faible: $F(56, 1016) = 1.48$, $p = .015$, $\eta^2 = .08$. Cet effet d'interaction très faible semble indiquer que la mesure des différents points n'ajoute pas d'information à une mesure plus globale de la F0. Afin de tester cette affirmation, la F0 moyenne de chaque expression a été régressée sur l'ensemble des points de F0 relevés. Les valeurs de F0 moyennes pour chaque expression ont été soumises préalablement à la soustraction du niveau moyen de F0 défini pour chaque locuteur, la régression a été effectuée simultanément sur les 9 points de F0 relevés. Une

ANOVA effectuée sur les valeurs résiduelles indique que l'émotion exprimée n'a plus d'effet sur les points de F0 relevés: $F(7, 1016) = 0.44$, $p = .879$, $\eta^2 = .00$; alors que l'effet d'interaction augmente très légèrement: $F(56, 1016) = 2.82$, $p < .001$, $\eta^2 = .13$. L'effet du point mesuré est évidemment préservé: $F(8, 1016) = 83.9$, $p < .001$, $\eta^2 = .40$.

Le graphique 2 représentent les moyennes et les écarts-types des résidus de la régression de la F0 moyenne des expressions pour chaque point de F0 relevé en fonction du type d'émotion exprimée. Ce graphique permet d'observer qu'après le contrôle de la F0 moyenne, des différences en fonction de l'émotion exprimée semblent apparaître pour le point '2max' et le point 'final'. Les moyennes et les écarts-types des résidus de la régression pour chaque émotion exprimée sont également représentés dans un tableau en annexe (E.3).

Graphique 2: Moyennes et écarts-types par type d'émotion exprimée pour les résidus de la régression de la F0 moyenne des expressions sur les points de F0 relevés



Des ANOVAs effectuées pour tester l'effet de l'émotion exprimée sur chacun de ces points confirment cette observation. Pour le deuxième maximum, $F(7, 125) = 6.91$, $p < .001$, $\eta^2 = .28$, le test post-hoc HSD de Tukey indique que la différence est due au fait que les moyennes des résidus de F0 pour la tristesse et l'anxiété sont plus faibles que la moyenne pour la colère froide, la colère chaude et la joie intense et que les moyennes pour la tristesse, l'anxiété et la joie calme sont plus faibles que les moyennes pour la colère chaude et la joie intense. Pour le point final: $F(7, 102) = 4.42$, $p < .001$, $\eta^2 = .23$, le test post-hoc HSD de Tukey indique que cette différence est due au fait que la moyenne pour les expressions de colère chaude est plus faible que les moyennes pour le désespoir, la peur panique, la tristesse, la joie calme et l'anxiété. D'après ce test, les moyennes pour la colère froide et la joie intense ne sont pas significativement différentes des autres moyennes.

6.2.2.3 Effet de l'émotion exprimée sur les pentes des contours stylisés

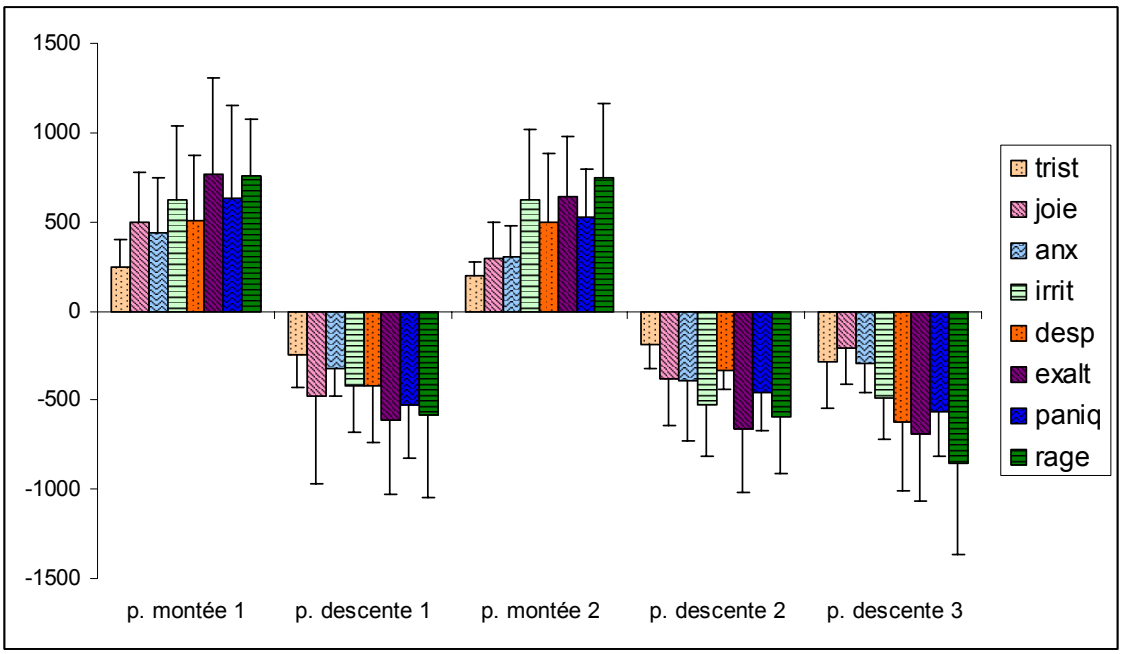
Les résultats des analyses présentées ci-dessus indiquent que les contours (la hauteur des points mesurés) pour différentes émotions exprimées sont essentiellement différenciés par le niveau global

de la F0. Les différences observées pour la hauteur du deuxième maximum et pour la hauteur du point final, après le contrôle de l'influence de la F0 moyenne des expressions, suggèrent d'autre part que l'amplitude des excursions locales de F0 (en particulier pour le deuxième "accent"), ainsi que la chute finale de la F0 sur la dernière syllabe pourraient être affectées par l'émotion exprimée. Afin d'examiner cette possibilité, des pentes ont été calculées pour les deux premières montées et pour les trois descentes codées.

> Le calcul des pentes pour les montées a été effectué en soustrayant le 1er minimum local (point '1min1' ou '2min1', en Hz) du maximum ('1max' ou respectivement '2 max', en Hz), puis en divisant la valeur obtenue par la durée (en secondes) de l'excursion entre le 1er minimum local et le maximum.
>
> Le calcul des pentes pour les descentes a été effectué en soustrayant le maximum local (point '1max', '2max' ou '3 max', en Hz) du 2ème minimum (respectivement '1min2', '2min2' ou 'final', en Hz), puis en divisant la valeur obtenue par la durée (en secondes) de l'excursion entre le maximum et le 2ème minimum local.

Le graphique 3 représente les moyennes et les écarts-types des pentes calculées en fonction du type d'émotion exprimée pour la première excursion de F0, pour la deuxième excursion de F0 et pour la descente finale. Ce graphique met en évidence des différences entre les émotions exprimées qui semblent assez similaires pour les cinq pentes codées. Ces pentes tendent à être plus fortes pour une partie des émotions qui comprennent une forte activation – en particulier pour la joie intense et la colère chaude – et plus faibles pour une partie des émotions qui comprennent une faible activation – en particulier la tristesse, la joie calme et l'anxiété. Les moyennes et les écarts-types des pentes sont également représentés dans un tableau en annexe (E.4) pour chaque émotion exprimée.

Graphique 3: Moyennes et écarts-type pour les pentes de F0 codées en fonction du type d'émotion exprimée



Des ANOVA effectuées pour tester l'effet de l'émotion exprimée sur chacune de ces pentes indiquent que l'émotion a un effet significatif sur la pente de la première montée ($F(7, 114) = 3.08$, $p = .005$, $\eta^2 = .16$), sur la pente de la deuxième montée ($F(7, 115) = 5.96$, $p < .001$, $\eta^2 = .27$, sur la pente de la deuxième descente ($F(7, 120) = 5.09$, $p < .001$, $\eta^2 = .23$) et sur la pente de la descente finale ($F(7, 101) = 6.01$, $p < .001$, $\eta^2 = .29$). L'effet de l'émotion n'est pas significatif pour la pente de la première descente ($F(7, 84) = 1.52$, $p = .172$, $\eta^2 = .11$).

La similarité des patterns observés pour les différentes émotions sur les 5 pentes suggère qu'une mesure plus générale de l'étendue de F0 pourrait rendre compte des différence observées sur l'ensemble des pentes. Afin de tester cette hypothèse, les ANOVAs présentées ci-dessus ont été répétées en utilisant l'étendue de F0 (définie comme la différence entre le minimum absolu de F0 et le maximum absolu de F0 de chaque contour) comme variable covariée. L'émotion exprimée n'a alors plus d'effet significatif sur la pente de la première montée ($F(7, 113) = 0.64$, $p = .722$, $\eta^2 = .04$). En revanche un effet tendanciel est préservé sur la pente de deuxième montée ($F(7, 114) = 2.07$, $p = .052$, $\eta^2 = .11$) et l'effet de l'émotion sur les pentes reste significatif après le contrôle de l'étendue globale des contours pour la deuxième descente ($F(7, 119) = 2.28$, $p = .033$, $\eta^2 = .12$) et la descente finale ($F(7, 100) = 2.73$, $p = .012$, $\eta^2 = .16$). Des comparaisons par paires ont permis de mettre en évidence les différences qui sont à l'origine de ces effets. Les différences significatives entre les pentes moyennes pour différentes émotions exprimées sont représentées dans le tableau 5.

Tableau 5: Différences entre les pentes moyennes par émotion exprimée pour la deuxième montée, la deuxième descente et la descente finale, après covariation de l'étendue de F0

deuxième montée			**deuxième descente**			**descente finale**		
trist	<	paniq, exalt, irrit, rage	trist, desp	<	anx, exalt	joie	<	paniq, desp, exalt, rage
joie	<	exalt, irrit, rage				anx	<	exalt, rage
anx	<	irrit, rage				trist, irrit, paniq	<	rage

La variabilité globale des contours de F0 (définie comme l'étendue de F0) n'explique donc pas toutes les différences observées entre les émotions exprimées en ce qui concerne les pentes des contours. L'effet de l'émotion exprimée sur les aspects des contours décrits ci-dessus est toutefois assez faible. Afin d'évaluer la possibilité de mettre en évidence d'autres différences significatives entre les émotions exprimées, trois aspects additionnels des contours stylisés ont été analysés. Il s'agit: de la position du maximum absolu de F0 dans chaque expression, des différences de hauteur entre les maxima locaux et de la "déclinaison" de la F0 du début à la fin de chaque expression.

6.2.2.4 Effet de l'émotion exprimée sur la position du point de F0 maximum

La position du maximum absolu de F0 a été définie pour chaque expression relativement aux trois segments identifiés lors de la stylisation des contours (premier "accent", deuxième "accent", syllabe finale) et relativement à la durée totale des expressions. Parmi les 144 expressions étudiées, trois expressions présentent des contours de F0 trop mal définis pour permettre d'identifier un véritable maximum. En conséquence, ces trois expressions (une expression de joie calme, une expression de tristesse et une expression d'anxiété) ne figurent pas dans les résultats présentés ci-dessous. Le tableau 4 indique, pour chaque émotion exprimée, la proportion (pourcentage) des maxima observés pour chacun des segments définis lors de la stylisation. Pour l'ensemble des expressions, le maximum absolu se situe le plus souvent sur le deuxième segment (acc2), 57% des maxima absolus sont observés sur ce segment. Les résultats présentés dans le tableau 6 indiquent que l'émotion exprimée semble influencer la position du maximum dans l'expression. On observe notamment que pour les émotions joie intense (exalt) et désespoir (desp), un tiers des maxima absolus se situent sur la syllabe finale contre respectivement 11% (i.e. 2 observations) et 6% (i.e. 1 observation) sur le premier segment; alors qu'on observe une tendance inverse pour les expressions de colère chaude (rage) et froide (irrit) qui atteignent leur valeur maximale de F0 sur le premier segment dans un tiers des cas contre respectivement 17% et 6% sur la syllabe finale. Les expressions de joie calme (joie) se distinguent tout particulièrement; le maximum de F0 pour ces expressions est atteint sur le premier segment dans 53% des cas, alors qu'aucune des 17 expressions de joie calme ne réalise son maximum de F0 sur la syllabe finale.

Tableau 6: Position du maximum de F0, proportion de la distribution sur les 3 segments distingués lors de la stylisation pour chaque émotion exprimée et pour la totalité des expressions.

	anx	joie	rage	irrit	paniq	trist	exalt	desp	total
acc1	24%	53%	33%	33%	17%	29%	11%	6%	26%
acc2	65%	47%	50%	61%	56%	65%	56%	61%	57%
final	12%	0%	17%	6%	28%	6%	33%	33%	17%
(N)	(17)	(17)	(18)	(18)	(18)	(17)	(18)	(18)	(141)

Bien que le nombre d'observations dans chaque cellule soit en principe trop faible pour l'utilisation de cette méthode, un chi^2 de Pearson a été calculé pour évaluer l'effet de l'émotion, ainsi que l'effet de la phrase et du locuteur sur la position du maximum. Les résultats de cette analyse sont donnés à titre indicatif. L'effet de l'émotion et de la phrases sont significatifs; $chi^2(14) = 24.38$, $p = .041$ pour l'émotion exprimée, $chi^2(2) = 22.99$, $p < .001$ pour la phrase. L'effet du locuteur sur la position du maximum est non significatif: $chi^2(16) = 13.02$, $p = .671$. La différence entre les deux phrases traitées est particulièrement remarquable: pour la première phrase ("hät san dig prong nju ven tsi"), le maximum de F0 est atteint dans 53% des cas sur le deuxième segment, contre 14% des cas sur le

premier segment et 3% des cas sur le segment final; alors que pour la deuxième phrase ("fi gött laich jean kill gos terr"), la distribution de la position du maximum est beaucoup plus équilibrée entre les différents segments; avec respectivement 22%, 28% et 21% des maxima sur le premier segment, le deuxième segment et le segment final.

Afin d'examiner plus en détail la question de la position du maximum absolu dans les expressions correspondant à différentes émotions, la position exacte du maximum de F0 a été définie relativement à la durée de chaque expression en divisant la position du maximum relativement au commencement de chaque expression par la durée totale de l'expression. Le graphique 4 représente les moyennes et les écarts-types de cette mesure pour chaque émotion exprimée et pour chaque phrase. Les valeurs correspondantes sont comprises dans le tableau 7 qui inclut également le nombre d'observations correspondant à chaque interaction phrase/émotion.

Graphique 4: Position du maximum de F0 relativement à la durée de chaque expression, moyennes et écarts-types par phrase et par émotion exprimée.

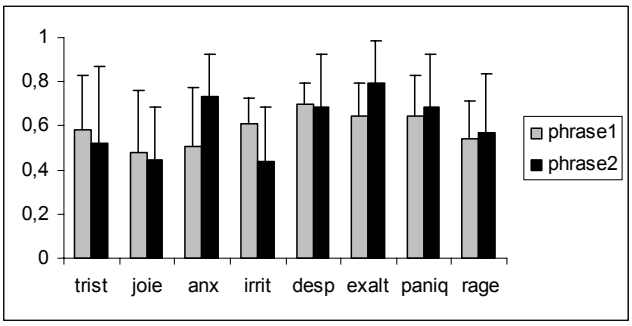


Tableau 7: Position du maximum de F0 relativement à la durée de chaque expression, moyennes et écarts-types par phrase et par émotion exprimée.

	trist	joie	anx	irrit	desp	exalt	paniq	rage
phrase1								
moy	0,58	0,48	0,51	0,61	0,70	0,64	0,64	0,54
sd	0,25	0,28	0,27	0,12	0,09	0,15	0,19	0,18
N	9	8	8	9	9	9	9	9
phrase2								
moy	0,52	0,45	0,73	0,44	0,69	0,80	0,68	0,57
sd	0,35	0,23	0,19	0,25	0,24	0,19	0,25	0,26
N	8	9	9	9	9	9	9	9

Le graphique 4 et le tableau 7 mettent en évidence des différences entre les émotions exprimées qui semblent un peu plus prononcées pour la deuxième phrase que pour la première phrase. On remarque par ailleurs qu'avec cette mesure continue de la position du maximum de F0 dans l'expression, les différences entre les émotions exprimées apparaissent moins tranchées qu'avec la mesure catégorielle présentée ci-dessus. Certaines observations effectuées à l'aide de la mesure catégorielle de la position du maximum sont toutefois renforcées par les résultats obtenus avec la mesure continue. Le contraste entre les expressions de joie calme (dont les maxima sont concentrés sur le 1er et le 2ème segment) et les expressions de joie intense (dont les maxima sont concentrés sur le 2ème segment et le segment final) est notamment confirmé par les résultats rapportés dans le graphique 4 et le tableau 7. Pour les expressions de joie calme, le point maximum du contour est atteint plus rapidement – en moyenne un peu avant le milieu des expressions – que pour les expressions de joie intense dont les contours atteignent leur maximum en moyenne à plus des 2/3 des expressions.

Les deux mesures (catégorielle et continue) de la position du maximum de F0 dans la phrase indiquent que la position du maximum est plus variable pour la deuxième phrase que pour la première phrase. La mesure continue de la position du maximum de F0 dans la phrase présente l'avantage de permettre d'évaluer simultanément l'effet de la phrase et de l'émotion et l'interaction de ces deux variables sur la position du maximum. Une ANOVA (2 x 8) effectuée pour tester l'effet de la phrase et de l'émotion révèle que seule l'émotion exprimée affecte significativement la position du maximum de F0 dans la phrase. D'après le test post-hoc de Tukey, l'effet principal de l'émotion ($F(7, 125) = 2.78$, $p = .010$, $\eta^2 = .14$) est dû à la différence entre la position moyenne du maximum de F0 pour les émotions de joie calme (à 46% de la durée totale des expressions) et la position moyenne du maximum de F0 pour les émotions de joie intense. (à 72% de la durée totale des expressions). La phrase ($F(1, 125) = 0.32$, $p = .574$, $\eta^2 = .00$) et l'interaction phrase/émotion ($F(7, 125) = 1.30$, $p = .255$, $\eta^2 = .07$) n'ont pas d'effet significatif sur la position du maximum.

6.2.2.5 Effet de l'émotion exprimée sur la hauteur relative des maxima locaux

La stylisation des contours a été effectuée de manière à définir trois maxima locaux pour chaque contour de F0 considéré. Les résultats présentés plus haut indiquent qu'il n'a pas été possible d'identifier trois maxima locaux pour toutes les expressions analysées. Toutefois, pour les expressions qui présentent plusieurs maxima locaux, il paraît intéressant d'examiner dans quelle mesure la hauteur relative de ces maxima est affectée par l'émotion exprimée; on peut en effet supposer qu'une différence importante entre la hauteur de deux maxima contigus pourrait se traduire par une saillance perceptive accrue pour le maximum le plus élevé. Dans la section précédente, il a été établi que le point de F0 le plus haut est le plus souvent atteint sur le deuxième segment défini pour la stylisation ('acc2'). La hauteur relative des maxima locaux a été en conséquence définie relativement au maximum réalisé sur le deuxième segment: la différence entre la hauteur du premier et du deuxième maximum définis pour la stylisation, ainsi que la différence entre la hauteur du deuxième et du troisième maximum ont été examinées. Le graphique 5 représente la moyenne et l'écart-type par émotion exprimée pour la différence de hauteur entre le deuxième maximum et le premier maximum (2max-1max) et pour la différence de hauteur entre le deuxième maximum et le troisième maximum (2max-3max). Le tableau 8 représente les valeurs correspondantes ainsi que le nombre d'observations par émotion exprimée (nombre d'expressions qui comprennent un premier et un deuxième maximum et nombre d'expressions qui comprennent un deuxième et un troisième maximum). Deux T-test (one-sample) ont été effectués pour chaque émotion exprimée afin d'évaluer statistiquement l'importance de la différence entre le deuxième et le premier maximum et

entre le deuxième et le troisième maximum. Les moyennes rapportées dans le tableau 8 sont suivies d'un astérisque lorsqu'elles sont statistiquement différentes de zéro.

Graphique 5: Différence entre la hauteur du 1er et du 2ème maximum et entre la hauteur du 2ème et du 3ème maximum, moyennes et écarts-types par émotion exprimée.

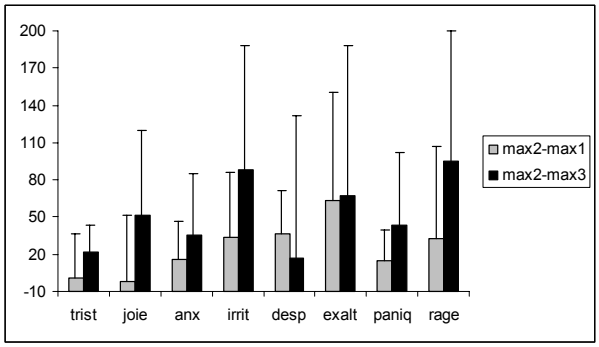


Tableau 8: Différence entre la hauteur du 1er et du 2ème max. et entre la hauteur du 2ème et du 3ème max., moyennes et écarts-types par émotion exprimée.

	trist	joie	anx	irrit	desp	exalt	paniq	rage
max2-max1								
moy	0,5	-1,8	15,8	33,7*	36,4*	63,5*	14,5*	32,2
sd	36,3	52,8	30,6	52,0	34,6	86,5	24,8	74,6
N	17	15	16	17	13	17	17	14
max2-max3								
moy	22,1*	51,8*	35,2*	87,9*	16,6	66,9	43,2*	94,5*
sd	21,7	68,1	49,8	100,1	115,0	121,2	59,0	105,0
N	10	15	12	15	15	15	15	15

* p<.05 (la moyenne est significativement différente de 0)

Les expressions de colère froide, de peur panique, de désespoir et de joie intense présentent un deuxième maximum significativement plus élevé que le premier maximum. D'autre part, les expressions de désespoir et de joie intense sont les seules expressions qui ne présentent pas de différence significative entre la hauteur du deuxième et du troisième maximum. On peut donc en déduire que "l'accentuation" des expressions de désespoir et de joie intense serait en moyenne plus marquée sur la deuxième excursion que sur la première excursion et que le niveau de F0 ne diminuerait pas sensiblement pour ces expressions avant la chute finale sur la dernière syllabe, alors que "l'accentuation" des expressions de colère froide et de peur panique serait également en moyenne plus marquée sur la deuxième excursion que sur la première excursion, mais que le niveau de F0 diminuerait sensiblement avant la chute finale sur la dernière syllabe pour ces expressions. Les expressions de tristesse, de joie calme, d'anxiété et de colère chaude ne seraient pas plus "accentuées" sur la deuxième excursion relativement à la première excursion et leurs niveaux de F0 diminueraient sensiblement avant la chute finale sur la dernière syllabe.

Deux ANOVAs effectuées pour évaluer l'effet de l'émotion exprimée sur la différence entre le premier et le deuxième maximum et sur la différence entre le deuxième et le troisième maximum révèlent un effet faible mais significatif de l'émotion ($F(7, 118) = 2.68$, $p = .013$, $\eta^2 = .14$) sur la différence entre le premier et le deuxième maximum. Selon le test post-hoc HSD de Tukey, cet effet est dû à la différence entre la moyenne pour la joie intense (la différence de hauteur entre le 1er accent et le deuxième accent est de 63 Hz en moyenne pour les expressions de joie intense) et la moyenne pour la joie calme et la tristesse (pour ces deux émotions la différence moyenne entre le premier et le deuxième accent est pratiquement nulle). Cette analyse permet de mettre en évidence que la différence observée ci-dessus relativement à la position du maximum de F0 dans la phrase pour les expressions de joie calme (le maximum de F0 est atteint relativement tôt dans le contour) et

dans les expressions de joie intense (le maximum de F0 est atteint relativement tard dans le contour) doit être interprétée avec précaution. Il est probable, en effet, que les expressions de joie intense tendent à se caractériser par une accentuation plus marquée sur la seconde moitié des expressions étudiées. Les expressions de joie calme, en revanche, ne sont pas nettement plus accentuées sur la première partie des phrases analysées, malgré que leur maximum de F0 soit atteint précocement. L'effet de l'émotion exprimée sur la différence entre le deuxième et le troisième maximum n'est pas significatif ($F(7, 104) = 1.46$, $p = .190$, $\eta^2 = .09$).

6.2.2.6 *Effet de l'émotion exprimée sur la "déclinaison" de la F0*

La "déclinaison" de la F0 est définie ici comme la différence (en Hz) entre la première valeur de F0 mesurée sur la première syllabe ('start') et la dernière valeur de F0 mesurée sur la syllabe finale ('final'), divisée par la durée (en secondes) qui sépare ces deux points. Le graphique 6 représente les moyennes et les écarts-types de cette déclinaison par émotion exprimée. Le tableau 9 représente les valeurs correspondantes, ainsi que le nombre d'observations comprises dans chaque moyenne.

Graphique 6: Déclinaison de la F0, moyennes et écarts-types par émotion exprimée.

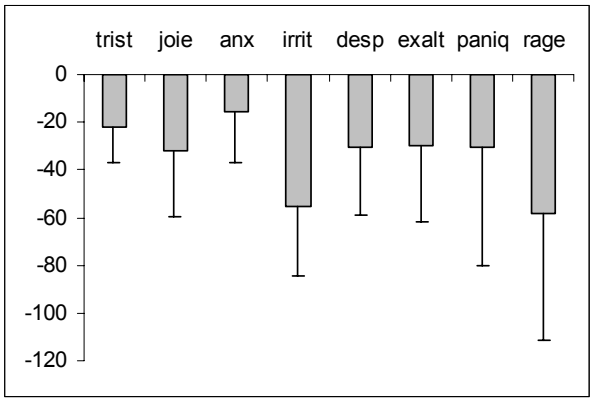


Tableau 9: Déclinaison de la F0, moyennes et écarts-types par émotion exprimée.

	trist	joie	anx	irrit	desp	exalt	paniq	rage
moy	-22,0	-31,8	-15,7	-55,4	-30,8	-29,5	-30,7	-58,1
sd	14,8	28,2	21,4	28,9	28,3	32,4	49,2	53,2
N	7	13	10	16	15	15	14	15

Une ANOVA effectuée pour tester l'effet de l'émotion exprimée sur la déclinaison de la F0 révèle un faible effet significatif ($F(7, 97) = 2.28$, $p = .034$, $\eta^2 = .14$). Le test post-hoc de Tukey en revanche n'identifie aucune paire d'émotions présentant des moyennes statistiquement différentes, seule une différence tendancielle apparaît entre les expressions d'anxiété qui diminuent de 16 Hz par seconde (en moyenne, du début à la fin de l'expression) et les expressions de colère chaude qui diminuent de 58 Hz par seconde (en moyenne, du début à la fin de l'expression).

L'émotion exprimée affecte plus fortement d'autre mesures. La pente ("déclinaison") entre le deuxième maximum et le point final (différence entre le point final et le point 2max en Hz, divisée par la distance en secondes entre ces deux points) est, par exemple, d'avantage affectée par l'émotion exprimée ($F(7, 96) = 5.18$, $p < .001$, $\eta^2 = .27$) que la "déclinaison" de la F0 du début à la

fin du contour telle que nous l'avons définie ci-dessus. En revanche cette mesure de la "déclinaison" entre le maximum situé au milieu des expressions et le point final se rapproche en réalité d'une mesure de l'étendue de la F0 qui peut être évaluée beaucoup plus simplement en calculant la différence entre le maximum absolu de F0 et le minimum absolu de F0. La corrélation observée entre ces deux mesures est de R = .68 (N = 104). L'effet de l'émotion exprimée sur la "déclinaison" entre le point '2max' et le point 'final' est fortement réduit et n'est plus statistiquement significatif lorsque l'étendue de F0 (la différence entre le minimum et le maximum de F0) est entrée dans l'analyse comme variable covariée ($F(7, 95) = 1.89$, $p = .080$, $\eta^2 = .12$).

Cette dernière observation peut être généralisée à d'autres mesures que l'on pourrait en principe dériver des contours stylisés. Des résumés plus simples des contours de F0 – tels que la F0 moyenne et l'étendue de F0 – permettent de rendre compte d'une large partie des différences observées pour différentes émotions exprimées. L'examen des contours reste toutefois intéressant dans la mesure où il donne accès aux caractéristiques plus spécifiques des contours qui sont à l'origine des effets observés au niveau des valeurs résumées de F0. L'effet, bien connu, de l'activation émotionnelle sur la F0 moyenne serait ainsi, d'après les résultats présentés ci-dessus, en grande partie imputable à un déplacement du niveau de base des contours et également, mais dans une moindre mesure, à des variations de l'étendue des excursions locales.

6.2.3 Conclusion

Pour l'ensemble des mesures dérivées des contours de F0 qui ont été décrites ci-dessus (hauteurs des points relevés, pentes des excursions, positions des maxima absolus, hauteurs relatives des maxima et "déclinaison"), la variabilité à l'intérieur des catégories correspondant aux émotions exprimées est très importante. L'effet de l'émotion exprimée sur ces mesures est en conséquence généralement assez faible. On remarque pourtant que malgré cette variabilité intra-émotionnelle très forte, certaines tendances semblent apparaître.

Les résultats exposés ci-dessus suggèrent notamment que pour certaines émotions – la colère chaude, la colère froide et la joie intense en particulier – la deuxième excursion de la F0 stylisée est plus important que pour d'autres émotions telles que la tristesse et la joie calme qui présentent une deuxième excursion nettement plus faible. Les analyses effectuées montrent également que cette différence ne peut être totalement expliquée par une différence de variabilité plus générale de la F0 pour les expressions correspondant à différentes émotions. De plus, la "forme" des contours semble également affectée par l'émotion exprimée. Un mouvement plutôt "uptrend" (terme emprunté à Ladd et al., 1985) et qui désigne ici une tendance à l'augmentation de la F0 au cours de l'expression et le maintien d'un niveau élevé jusqu'à la chute finale) a par exemple été observé pour les

expressions de désespoir et de joie intense, alors que les expressions de tristesse et de joie calme semblent présenter un mouvement plutôt "downtrend" (un maximum de F0 précoce suivit une diminution progressive jusqu'à la chute finale) plus conforme à ce que l'on attendrait dans le cadre d'une production de parole normale (non émotionnelle). La "chute finale" est également affectée par l'émotion exprimée. Les expressions de colère chaude et de joie intense présentent notamment des diminutions plus importantes de la F0 dans la dernière syllabe que les expressions d'anxiété et de joie calme.

6.3 Perception de l'émotion dans la synthèse vocale

Bien que des mesures plus générales de la F0, telles que la F0 moyenne ou l'étendue de F0 parviennent à rendre compte d'une large partie de l'effet de l'émotion exprimée sur les contours de F0, il semble donc que des caractéristiques du contour de F0 – telles que le niveau de base, la proéminence relative de différentes excursions ou encore la chute finale du contour – sont affectées par l'émotion exprimée. Afin d'évaluer dans quelle mesure des caractéristiques aussi simples des contours parviennent à communiquer une information émotionnelle en l'absence d'autres indices vocaux, différents contours de F0 ont été appliqués sur des expressions produites par un système de synthèse Text-To-Speech (TTS). La qualité émotionnelle perçue de ces expressions a été ensuite évaluée dans des tests de perceptions. Les procédures utilisées et les résultats obtenus sont décrits ci-dessous.

6.3.1 Synthèse des expressions

Manipulation systématique des contours de F0 (synthèse TTS)

Les deux séquences de syllabes qui composent les expressions émotionnelles décrites ci-dessus ont été utilisées. Une stylisation simple des contours de F0 (v. figure 2) a été définie de manière à produire deux "accents" (a1 et a2) perceptibles et qui donnent une impression naturelle pour ces séquences de syllabes. Ce mouvement de F0 a été appliqué sur des expressions produites par le système de synthèse KALI avec deux voix différentes (une voix masculine et une voix féminine). Des variations systématiques de la hauteur du niveau de base (bas/haut), de la taille des deux excursions locales (a1 et a2, faibles/fortes) et du mouvement final de la F0 sur la dernière syllabe (descendant/plat/montant) ont été effectuées. Au total, 96 expressions ont été créées : 2 séquences de syllabes * 2 voix * 2 niveaux de base * 2 accents * 2 niveaux d'accents * 3 mouvements finaux. La différence entre les deux niveaux de base est de 4 tons. La différence entre un accent faible et un accent fort est de 3 tons. Le mouvement final plat se termine sur le niveau de base, le mouvement

montant se termine à 4 tons au-dessus du niveau de base, le mouvement final descendant se termine à 4 tons au-dessous du niveau de base.

Des précisions techniques sur le synthétiseur KALI et sur les manipulations de la parole effectuées sont disponibles dans Morel & Lacheret-Dujour (2001) et dans Morel & Bänziger (2004). Afin d'illustrer les différences de F0 introduites dans les expressions de synthèse, des contours de F0 extraits de trois expressions produites par le synthétiseur sont superposés dans la figure 3. Le contour représenté par une ligne continue noire correspond à une expression produite avec la voix de synthèse féminine, avec un niveau de F0 'bas', deux accents (a1 et a2) 'faibles' et un mouvement final 'descendant'. Le contour représenté par une ligne rouge discontinue correspond à une expression produite avec les mêmes diphones, avec le même niveau de F0 (niveau 'bas'), mais deux accents (a1 et a2) 'forts' et un mouvement final 'plat'. Le contour représenté par une ligne discontinue bleue est identique au contour représenté par la ligne discontinue rouge, à l'exception de la chute finale qui est 'montante' pour l'expression dont ce contour a été extrait.

Figure 2: Stylisation de la F0 appliquée aux séquences de syllabes produites à l'aide du système de synthèse KALI

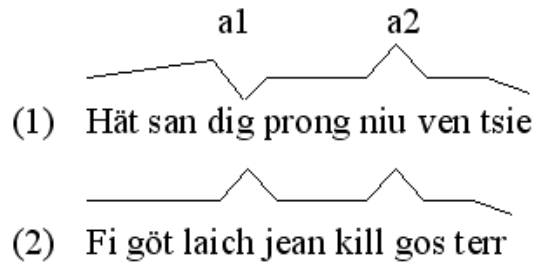
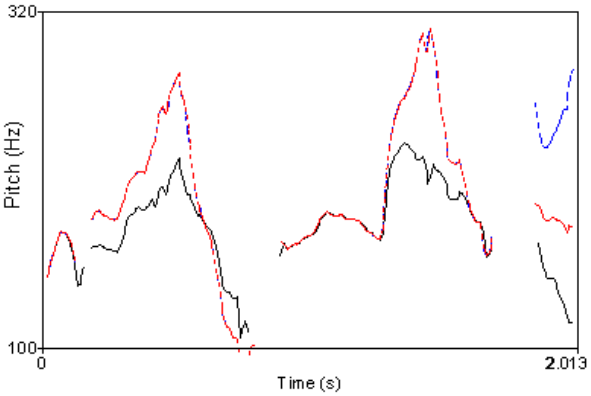


Figure 3: Contours de F0 pour différentes expressions crées avec le système de synthèse



a1 et a2 faibles (ligne noire continue), a1 et a2 forts (ligne rouge discontinue); final montant (bleu), plat (rouge), descendant (noir).

Intonation copiée (synthèse TTS et resynthèse)

A titre de comparaison avec les expressions de synthèse décrites ci-dessus, 16 expressions supplémentaires ont été produites avec le système de synthèse utilisé pour la manipulation systématique des contours (TTS - KALI). Ces 16 expressions ont été réalisées en transposant, sur la voix de synthèse, l'intonation (la F0, l'intensité et la durée) produite par des acteurs exprimant des émotions. A cette fin, 16 expressions émotionnelles (8 types d'émotions * 2 "phrases") ont été sélectionnées parmi les 144 expressions dont les contours ont été analysés ci-dessus (section 5.2). Des expressions dont la qualité émotionnelle a été bien reconnue lors du test perceptif décrit à la section 3 et dont la qualité acoustique a permis une relativement bonne définition du contour de F0

ont été choisies. Les deux "phrases" (1. "hät san dig prong nju ven tsi", 2. "fi götte laich jean kill gos terr") sont donc prononcés par différents acteurs/trices et communiquent 8 types d'émotions : colère chaude ('rage') et colère froide ('irrit'), anxiété ('anx') et peur panique ('paniq'), tristesse ('trist') et désespoir ('desp'), joie calme ('joie') et joie intense ('exalt'). Le tableau 10 représente les 16 expressions sélectionnées avec un ensemble de caractéristiques incluant: l'émotion exprimée, l'identificateur de l'acteur/trice qui prononce l'expression, la phrase prononcée et l'intensité moyenne de peur, joie, colère et tristesse perçue pour chaque expression lors du test perceptif.

Tableau 10: Caractéristiques des expressions sélectionnées pour le transfert de l'intonation sur la voix de synthèse

expression	emotion exprimée	phrase	acteur/ actrice	intensité perçue de peur	joie	tristesse	colère
A11103	anx	1	3	**5,7**	2,0	4,0	1,7
A22103	anx	2	3	**7,8**	1,2	3,9	1,9
P11210	paniq	1	10	**8,2**	2,7	5,0	6,4
P12203	paniq	2	3	**8,5**	4,3	2,1	3,0
F21210	joie	1	10	0,8	**7,0**	1,1	0,8
F22211	joie	2	11	0,4	**6,2**	0,9	0,1
U11107	exalt	1	7	1,4	**8,1**	2,4	2,2
U22210	exalt	2	10	0,4	**8,2**	0,7	0,3
T21209	trist	1	9	5,1	1,1	**8,9**	1,7
T22203	trist	2	3	5,4	1,3	**8,9**	0,7
Z21204	desp	1	4	3,2	1,2	**7,1**	1,4
Z12212	desp	2	12	3,9	1,8	**8,1**	1,6
K11103	irrit	1	3	1,0	0,8	0,8	**6,8**
K12209	irrit	2	9	0,8	1,4	1,6	**7,0**
H11108	rage	1	8	1,6	3,4	1,5	**9,1**
H22109	rage	2	9	1,1	0,6	0,8	**9,4**

L'intonation des expressions présentées dans le tableau 8 a été copiée sur 16 expressions produites à l'aide du synthétiseur KALI. D'autre part, les 16 expressions originales ont été resynthétisées avec l'intonation neutre produite par le système de synthèse KALI pour les "phrases" (séquences de syllabes) considérées. La figure 4 représente un contour de F0 extrait de l'une des expressions originales (l'expression de joie calme produite par l'actrice 10); la figure 5 représente le contour de F0 extrait de l'expression de synthèse (KALI) sur laquelle l'intonation de cette expression originale a été copiée; la figure 6 représente le contour de F0 de l'expression originale resynthétisée avec l'intonation "neutre" du système de synthèse.

Les figures 4 à 6 illustrent "la copie croisée" de l'intonation qui a été opérée pour les 16 expressions sélectionnées. Au total, 48 expressions - 16 expressions originales, 16 expressions des synthèse avec l'intonation émotionnelle originale (et la qualité vocale "neutre" du synthétiseur), 16 expressions resynthétisées avec une intonation "neutralisée" (et une qualité vocale émotionnelle)

ont été ajoutées aux 96 expressions dont les contours ont été systématiquement modifiés dans les études de jugements présentées ci-dessous.

Figure 4: Contour de F0 extrait d'une expression de joie calme produite par l'actrice 10

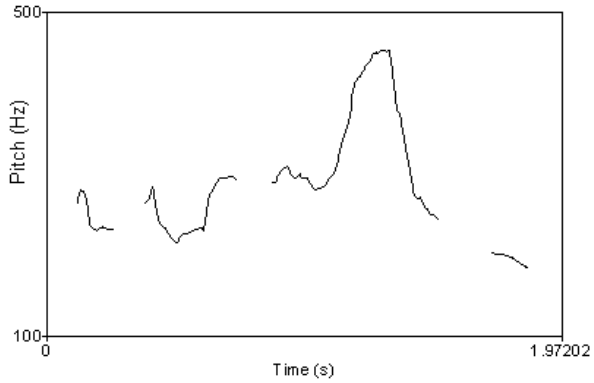


Figure 5: Contour de F0 extrait de l'expression de synthèse sur laquelle l'intonation émotionnelle produite par l'actrice a été copiée

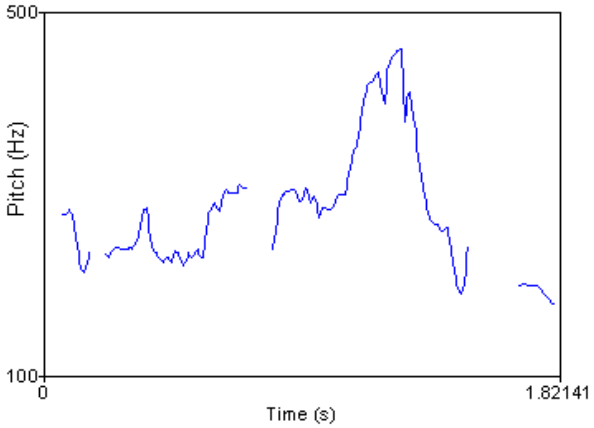
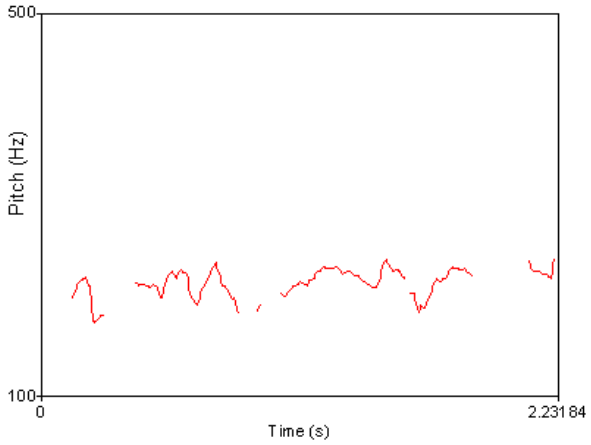


Figure 6: Contour de F0 extrait après la resynthèse de l'expression avec l'intonation "neutre" du système de synthèse



6.3.2 Première évaluation perceptive

6.3.2.1 Procédure

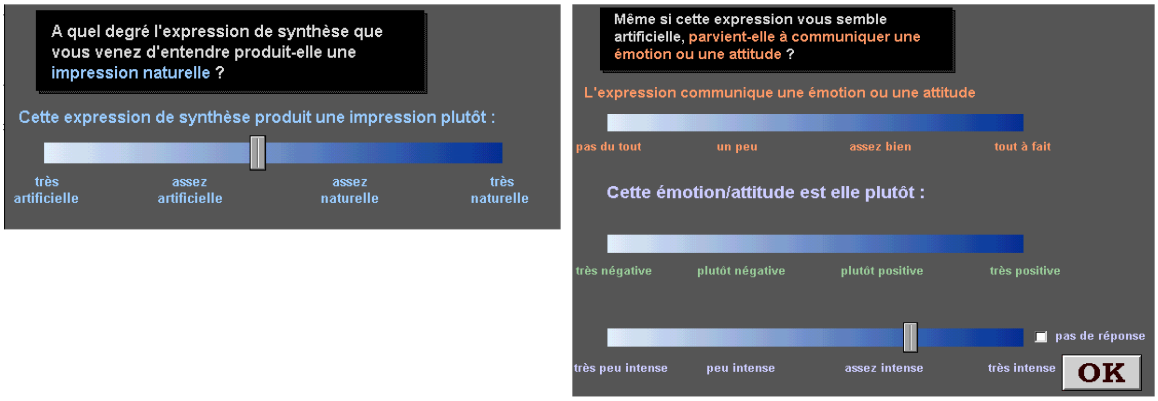
Une première évaluation de la qualité émotionnelle et, également, du naturel des expressions a été obtenue en soumettant l'ensemble des expressions à l'appréciation d'un groupe de 7 auditeurs (1 homme et 6 femmes; âge moyen = 32 ans, écart-type = 3.6). Les 16 expressions émotionnelles sélectionnées, les 32 expressions produites par la copie croisée de l'intonation entre les expressions émotionnelles originales et le système de synthèse et les 96 expressions de synthèse dont le contour de F0 a été systématiquement modifié (soit au total 144 expressions) ont été présentées durant la même session dans un ordre aléatoire différent pour chaque auditeur. Les auditeurs ont évalué pour chaque expression présentée:

- A quel degré l'expression produit une impression naturelle versus artificielle ('naturel').

- A quel degré l'expression parvient à communiquer une émotion ou une attitude ('émotionnel').
- A quel point cette émotion/attitude est positive versus négative ('valence').
- Quel est le degré d'intensité de cette émotion/attitude ('intensité')

Les jugements ont été effectués sur des échelles visuelles analogues, les réponses ont été enregistrées sur des échelles continues de 0 à 10. La formulation exacte des questions posées et la forme des échelles de réponses présentées aux participants sont représentées sur la figure 7.

Figure 7: Reproduction des questions et des échelles de réponses utilisées pour la première évaluation perceptive des expressions



6.3.2.2 *Résultats*

Les corrélations intraclasses des réponses données par les 7 auditeurs ayant participé à cette première étude de jugement sont représentées dans le tableau 11. Ces indices de fidélité des réponses ont été calculés relativement aux réponses données pour l'ensemble des 144 expressions évaluées et relativement aux réponses données seulement pour les 96 expressions dont les contours de F0 ont été systématiquement modifiés. Les indices présentés dans ce tableau indiquent que les réponses de différents auditeurs ne sont pas fidèles pour les expressions de synthèse dont le contour a été systématiquement modifié. Si l'on considère la totalité des expressions évaluées (les expressions naturelles, les expressions produites par la copie croisée de l'intonation et les expressions dont le contour de F0 a été modifié systématiquement), la fidélité des réponses des 7 auditeurs est très élevée (R = .93 à .94). En revanche, si l'on considère uniquement les réponses données pour les contours de F0 systématiquement modifiés, l'indice de fidélité chute à R = .11 pour le jugement relatif au degré d'émotion/attitude exprimée, cet indice (R = .11) correspond dans le cas présent à une corrélation moyenne nulle (r = .02) entre les réponses des différents auditeurs.

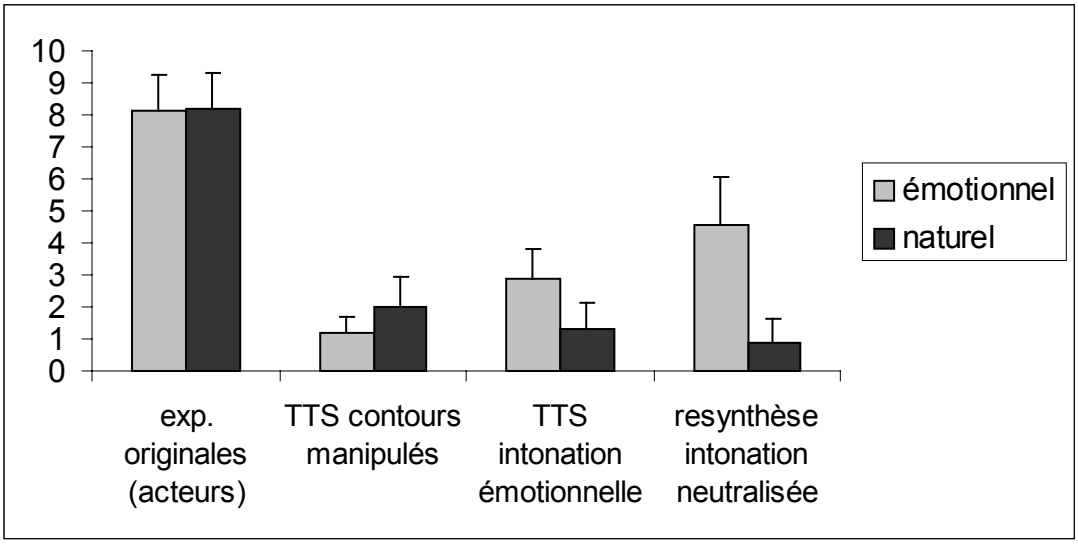
Tableau 11: Corrélations intraclasse (fidélité inter-juges) pour la totalité des jugements et pour les jugements relatifs aux contours de F0 systématiquement modifiés

Expressions	Naturel		Emotionnel		Valence		Intensité	
	r	R	r	R	r	R	r	R
Totalité	.71	.94	.67	.94	.66	.93	.64	.93
TTS contours manipulés	.22	.66	.02	.11	.11	.47	.04	.22

r = single measure intraclass correlation (N = 7) R = average measure intraclass correlation (N = 7)

Dans cette étude de jugements, 4 catégories d'expressions très différentes ont été évaluées par les auditeurs durant la même session. Les jugements qu'ils ont effectués sont, en conséquence, avant tout fonction du type d'expression présenté. Le graphique 7 représente les moyennes des jugements de la qualité naturelle et émotionnelle pour (1) les 16 expressions originales (produites par les acteurs), pour (2) les 96 expressions de synthèse avec les contours systématiquement manipulés, pour (3) les 16 expressions de synthèse avec l'intonation des expressions originales et pour (4) les 16 expressions resynthétisées dont l'intonation à été neutralisée. Les écarts-types rapportés dans ce graphique représentent la variabilité due aux différentes expressions inclues dans chaque moyenne, la variabilité introduite par les différents juges a été retirée en calculant d'abord un jugement moyen pour chaque expression.

Graphique 7: Moyenne et écarts-types des jugements concernant la qualité naturelle et la qualité émotionnelle pour les 4 catégories d'expressions évaluées



Le graphique 7 montre que les jugements de la qualité naturelle et de la qualité émotionnelle différencient essentiellement les expressions originales ("naturelles", i.e. produites par des acteurs) des expressions créées avec le synthétiseur TTS et des expressions resynthétisées. Les expressions produites par les acteurs sont jugées très naturelles et très émotionnelles, alors que les expressions manipulées sont jugées globalement beaucoup moins naturelles et beaucoup moins émotionnelles.

Deux ANOVAs effectuées pour tester l'effet du type d'expression sur les jugements relatifs à la qualité naturelle ($F(3, 140) = 248.95$, $p < .001$, $\eta^2 = .84$) et sur les jugements relatifs à la qualité émotionnelle ($F(3, 140) = 381.05$, $p < .001$, $\eta^2 = .89$) confirme la présence d'un effet très important du type d'expressions sur ces deux jugements. Comme mentionné ci-dessus, ces effets sont en grande partie dus à la différence entre les jugements pour les expressions "naturelles" d'une part et pour les expressions manipulées (TTS et resynthèse) d'autre part. Toutefois, le test post-hoc HSD de Tukey indique que d'autres différences entre les catégories d'expressions contribuent également à ces effets. Les expressions TTS dont les contours ont été systématiquement manipulés sont jugés comme significativement ($p<.001$) moins émotionnelles que les deux autres catégories d'expressions manipulées. Les expressions resynthétisées pour "neutraliser" l'intonation sont quant à elles jugées comme significativement ($p<.001$) plus émotionnelles que les deux autres catégories d'expressions manipulées. Les expressions qui conservent la qualité vocale originale mais dont l'intonation est "neutralisée" par la resynthèse sont donc évaluées comme plus émotionnelles que les expressions qui ont été produites de manière à présenter différents contours intonatifs (manipulations systématiques de la F0 et reproduction de l'intonation produite par les acteurs). Cette observation est d'autant plus intéressante qu'elle n'est pas attribuable à une perception plus naturelle des expressions produites par la resynthèse. Les expressions produite par la resynthèse sont jugées significativement ($p<.001$) moins naturelles que les expressions dont les contours ont été systématiquement manipulés (le degré de naturel perçu n'est en revanche pas significativement différent pour les expressions produites par la resynthèse et les expressions TTS qui reproduisent l'intonation émotionnelle).

Les jugements très faibles relatifs à la qualité émotionnelle perçue pour les expressions de synthèse limitent la variabilité des réponses pour les questions relatives à l'intensité et à la valence de l'émotion communiquée. Les réponses données par les auditeurs sont le plus souvent nulles pour ces questions (l'option 'pas de réponse' est sélectionnée) lorsqu'elles sont posées relativement aux expressions dont les contours de F0 ont été systématiquement modifiés.

Les résultats présentés ci-dessus semblent indiquer que les expressions qui présentent uniquement des variations de l'intonation – en particulier les expressions dont les contours de F0 ont été systématiquement modifiés – ne sont pas perçues comme pouvant communiquer une émotion. Les jugements relatifs à la qualité émotionnelle de ces expressions sont très faibles et ne sont pas fidèles. Il reste toutefois possible que l'absence de qualité émotionnelle perçue et l'absence de fidélité inter-juges soient, en partie, dues à un effet de contraste entre les expressions dont les contours de F0 ont été manipulés et les expressions "naturelles" produites par les acteurs. Les expressions dont le contour de F0 a été modifié seraient évaluées globalement comme peu

naturelles et peu émotionnelles en comparaison avec les expressions naturelles; et une potentielle différenciation entre les expressions dont les contours de F0 ont été modifiés serait atténuée dans ce contexte. Afin de mettre en évidence des différences éventuelles entre des expressions appartenant à une même catégorie – particulièrement pour les expressions dont les contours de F0 ont été manipulés – un deuxième test perceptif a été mis en place.

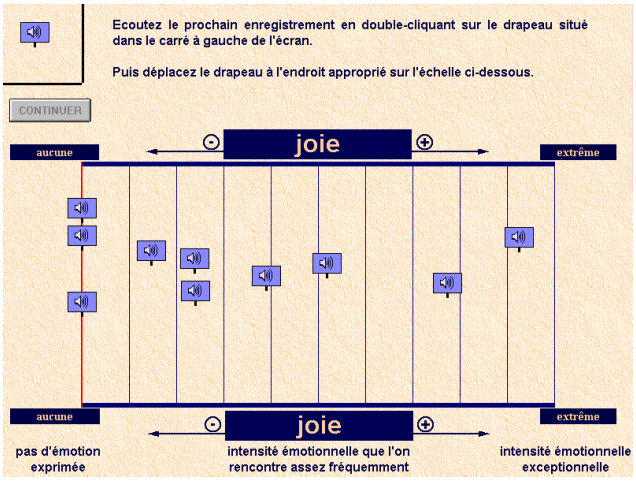
6.3.3 Seconde évaluation perceptive

La seconde évaluation perceptive des expressions de synthèse décrite ci-dessous a été effectuée avec deux objectifs. Premièrement, la méthode utilisée a été modifiée dans le but d'améliorer la précision des réponses données par les auditeurs. Deuxièmement, les questions posées aux auditeurs ont été modifiées de manière à cibler directement l'intensité émotionnelle communiquée par les expressions de synthèse pour des émotions spécifiques.

6.3.3.1 Procédure

Quatre dimensions émotionnelles représentant l'intensité de colère, de joie, de tristesse et de peur perçues (la figure 8 reproduit l'échelle utilisée pour évaluer l'intensité de joie perçue) ont été successivement présentées à des auditeurs qui avaient pour tâche de placer sur ces échelles des icônes représentant les expressions. Dans cette procédure, les auditeurs sont libres de réécouter les expressions aussi souvent qu'ils le désirent en cliquant sur les icônes qui représentent les expressions. Ils (dé)placent sur chaque dimension émotionnelle évaluée les expressions/icônes qu'ils peuvent comparer directement. Cette méthode augmente la précision des réponses données, relativement à la méthode traditionnelle dans laquelle les expressions sont présentées successivement aux auditeurs qui doivent donner une évaluation sur une échelle immédiatement après la présentation de chaque expression sans point de référence externe.

Figure 8: Echelle présentée aux auditeurs pour l'évaluation de l'intensité de joie perçue dans les expressions de synthèse



Quinze auditeurs (13 femmes et 2 hommes; âge moyen = 26 ans, écart-type = 6.2) ont évalué d'abord les expressions de synthèse dont la F0 a été systématiquement manipulée à l'aide de cette procédure. En raison du temps nécessaire pour réaliser ce type de jugements, seules les 48 expressions produites avec la voix de synthèse féminine ont été utilisées. Elles ont été présentées en deux blocs de 24 expressions correspondant aux deux énoncés utilisés. L'ordre de présentation de ces deux blocs d'expressions a été défini aléatoirement pour chaque auditeur. Les 4 échelles émotionnelles ont été présentées successivement dans un ordre aléatoire défini pour chaque bloc d'expressions et pour chaque auditeur. Après une courte pause, les mêmes auditeurs ont évalué les 16 expressions qui ont été produites en copiant l'intonation des 16 expressions naturelles sélectionnées sur la voix de synthèse, ainsi que les 16 expressions qui ont été produites en imposant l'intonation du système de synthèse sur les 16 expressions naturelles resynthétisées. L'ordre de présentation de ces deux ensembles d'expressions a été défini aléatoirement pour chaque auditeur.

6.3.3.2 Résultats

Fidélité inter-juges

Cette procédure n'a permis d'améliorer que partiellement la fidélité inter-auditeurs des évaluations pour les expressions produites en manipulant systématiquement les contours de F0. Les évaluations de l'intensité de joie et de peur fournies par les 15 auditeurs pour ces expressions présentent des corrélations relativement élevées alors que les jugements des 15 auditeurs concernant l'intensité de colère perçue et l'intensité de tristesse perçue ne sont pas corrélés. Les indices de fidélité sont représentés dans le tableau 12 pour les jugements de joie, peur, colère et tristesse. Ils ont été calculés séparément pour les 48 expressions dont le contour de F0 a été systématiquement modifié, pour les 16 expressions TTS qui reproduisent l'intonation des expressions émotionnelles "naturelles" et pour les 16 expressions dont l'intonation a été "neutralisée" par la resynthèse.

Pour les expressions produites par la copie croisée de l'intonation (expressions TTS avec l'intonation émotionnelle et expressions resynthétisées avec la prosodie "neutralisée"), les jugements des auditeurs sont relativement bien corrélés pour les quatre qualités émotionnelles évaluées.

Tableau 12: Corrélations intraclasse (fidélité inter-juges) pour les jugements relatifs aux 3 catégories d'expressions jugées successivement

Expressions	joie		peur		colère		tristesse	
	r	R	r	R	r	R	r	R
TTS contours manipulés	.35	.89	.61	.96	.04	.36	.05	.42
TTS intonation émotionnelle	.25	.83	.24	.82	.32	.88	.33	.88
Resynth. intonation neutralisée	.41	.91	.36	.90	.40	.91	.39	.91

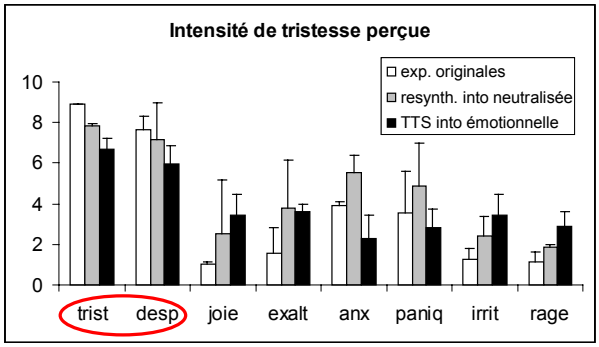
r = single measure intraclass correlation R = average measure intraclass correlation ; N = 15

Les résultats obtenus sont présentés ci-dessous en deux parties, en premier lieu pour les expressions produites par la copie croisée, puis pour les expressions dont les contours de F0 ont été manipulés.

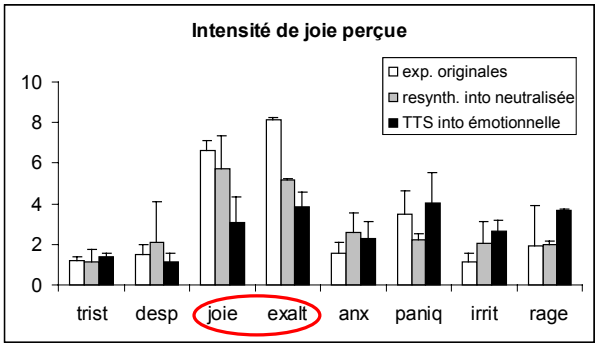
Expressions produites par la copie croisée de l'intonation

Les graphiques 8 à 11 représentent les moyennes et les écarts-types de l'intensité émotionnelle perçue par émotion exprimée pour les 16 expressions émotionnelles produites par les acteurs (barres blanches), pour les 16 expressions dont l'intonation a été "neutralisée" par la resynthèse (barres grises) et pour les 16 expressions TTS qui reproduisent l'intonation des expressions d'origine (barres noires). Au préalable, une moyenne de l'intensité perçue de tristesse, de joie, de peur et de colère a été calculée pour chacune des expressions de chaque catégorie. En ce qui concerne les 16 expressions produites par la resynthèse et les 16 expressions produites par le synthétiseur TTS, ces moyennes ont été calculées à partir des jugements effectués par les 15 auditeurs recrutés pour cette étude. En ce qui concerne les expressions émotionnelles produites par les acteurs, ces moyennes sont représentées dans le tableau 10 et sont basées sur des jugements effectués par 16 auditeurs dans une étude antérieure. Les écart-types représentés dans les graphiques 8 à 11 correspondent donc à la variabilité introduite par les deux expressions représentant chaque émotion exprimée, la variabilité due aux différents auditeurs n'est pas représentée sur ces graphiques.

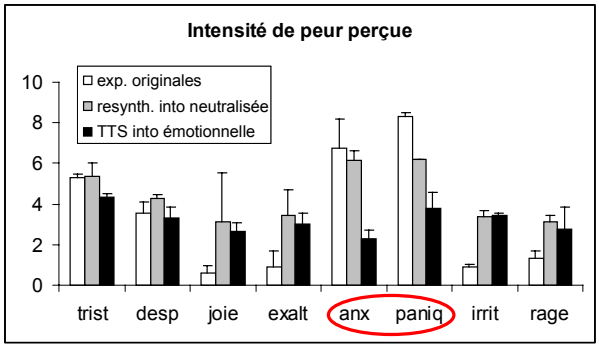
Graphique 8: Intensité de tristesse perçue par émotion exprimée pour 3 catégories d'expressions



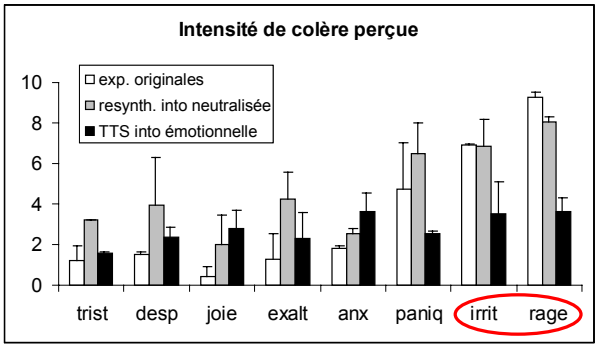
Graphique 9: Intensité de joie perçue par émotion exprimée pour 3 catégories d'expressions



Graphique 10: Intensité de peur perçue par émotion exprimée pour 3 catégories d'expressions



Graphique 11: Intensité de colère perçue par émotion exprimée pour 3 catégories d'expressions



L'examen des valeurs correspondant aux émotions exprimées qui sont encerclées sur les graphiques 8-11 indiquent que la perception de l'émotion exprimée est dégradée dans une plus forte mesure

pour les expressions produites par le système de synthèse avec l'intonation émotionnelle (barres noires) que pour les expressions dont l'intonation a été "neutralisée" par la resynthèse (barres grises). Seule la tristesse semble être encore assez bien communiquée dans les expressions qui ont été produites en copiant l'intonation des expressions naturelles sur la voix de synthèse; alors que pour les expressions produites en "neutralisant" l'intonation par la resynthèse, les émotions exprimées restent relativement bien reconnues; seules les deux expressions de joie intense (exalt) reçoivent des attributions d'intensité de joie qui apparaissent nettement plus faibles que les attributions pour les expressions originales lorsque l'intonation est "neutralisée" par la resynthèse.

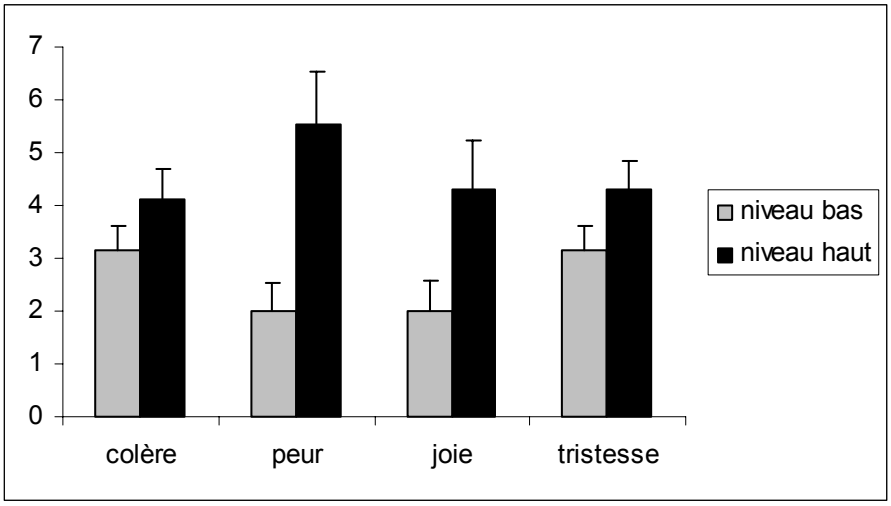
Des ANOVAs (3 catégories d'expressions x 8 types d'émotions exprimées) effectuées sur les jugements moyens calculés pour chaque expression indiquent que seule l'émotion exprimée affecte significativement l'intensité moyenne de tristesse perçue ($F(7, 24) = 19.75$, $p < .001$, $\eta^2 = .85$). Selon le test post-hoc de Tukey, cet effet est dû au fait que les moyennes pour les expressions de tristesse et de désespoir sont perçues comme exprimant une intensité de tristesse plus importante que les autres expressions, indépendamment des 3 catégories d'expression. En ce qui concerne les jugements relatifs à l'intensité de joie perçue, l'effet principal de l'émotion exprimée ($F(7, 24) = 18.69$, $p < .001$, $\eta^2 = .84$) et l'effet d'interaction entre l'émotion exprimée et la catégorie des expressions ($F(14, 24) = 3.49$, $p = .003$, $\eta^2 = .67$) sont significatifs. Pour l'intensité de peur perçue les 2 effets principaux (émotion exprimée $F(7, 24) = 23.56$, $p < .001$, $\eta^2 = .87$; catégorie des expressions $F(2, 24) = 10.13$, $p = .001$, $\eta^2 = .46$) et l'effet d'interaction ($F(14, 24) = 7.16$, $p < .001$, $\eta^2 = .81$) sont significatifs. Pour l'intensité de colère perçue les 2 effets principaux (émotion exprimée $F(7, 24) = 18.91$, $p < .001$, $\eta^2 = .85$; catégorie des expressions $F(2, 24) = 12.40$, $p < .001$, $\eta^2 = .51$) et l'effet d'interaction ($F(14, 24) = 4.05$, $p = .001$, $\eta^2 = .70$) sont également significatifs. L'examen des valeurs représentées dans les graphiques 9, 10 et 11 suggère que les effets d'interaction significatifs pour l'intensité de joie, de peur et de colère perçues reflètent le fait que les émotions exprimées qui correspondent à la qualité émotionnelle évaluée (encerclées sur les graphiques) reçoivent des jugements plus élevés que les autres émotions exprimées pour les expressions originales et les expressions produite par la resynthèse (intonation "neutralisée"), alors que les expressions TTS qui reproduisent l'intonation des expressions émotionnelles ne présentent pas ce pattern.

Manipulation des contours des expressions TTS

En ce qui concerne les expressions dont les contours de F0 ont été systématiquement modifiés, l'effet principal observé est dû à la manipulation du niveau de base de la F0 (haut versus bas) qui affecte assez fortement les jugements émotionnels (v. graphique 12). La direction de cet effet est la

même pour les 4 types d'émotions évaluées (colère, peur, joie, tristesse); l'intensité émotionnelle perçue est plus forte lorsque le niveau de F0 est 'haut' et plus faible lorsque le niveau de F0 est 'bas'. Toutefois, la magnitude de cet effet dépend du type d'émotion évalué. L'intensité de peur perçue est plus particulièrement affectée par le niveau de la F0, alors que l'influence du niveau de la F0 sur les jugement de colère et de tristesse est beaucoup plus limitée.

Graphique 12: Intensité de colère, de peur, de joie et de tristesse perçues en fonction du niveau de la F0 imposé sur les expressions de synthèse (écarts-types calculés pour 24 expressions avec un niveau bas et 24 expressions avec un niveau haut)



La fidélité relativement bonne des jugements obtenus de la part des 15 auditeurs pour les jugements de peur et de joie (v. tableau 12) est essentiellement due à la constance avec laquelle les auditeurs ont utilisé le niveau de base de la F0 pour effectuer ces deux types de jugements. Pour les jugements de tristesse et de colère en revanche la fidélité des jugements est très faible. Les évaluations moyennes obtenus pour chaque expression sur ces deux échelles émotionnelles ne sont dès lors pas fiables. Afin de tester malgré tout la présence de différences attribuables aux manipulations des contours effectuée, des ANOVAS à mesures répétées (facteurs intra/within: 2 phrases x 2 niveaux de base x 2 hauteurs de la première excursion x 2 hauteurs de la deuxième excursion x 3 mouvements finaux) ont été effectuées. Le tableau 13 représente la totalité des effets principaux et des effets d'interactions significatifs qui sont apparus pour les 4 jugements émotionnels.

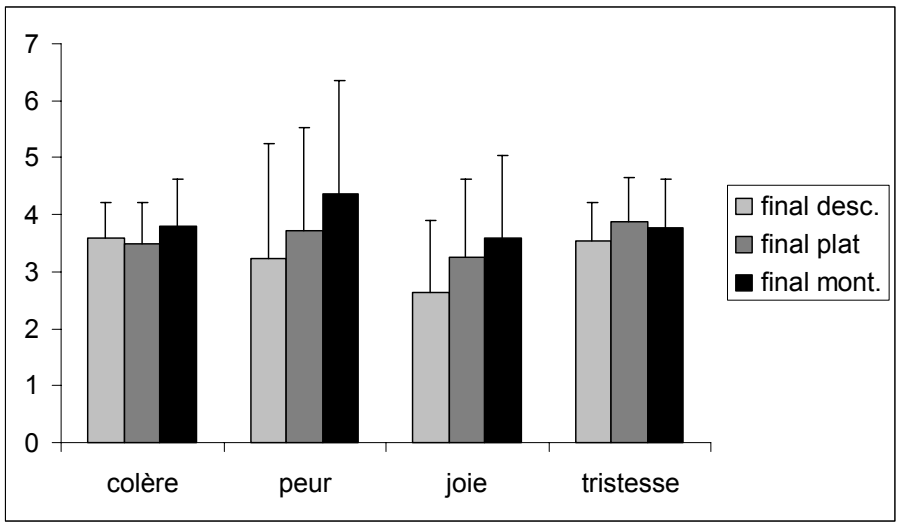
Seuls 2 effets sur 31 sont significatifs pour l'évaluation de l'intensité de tristesse, alors qu'avec une probabilité d'erreur du premier type de 5%, 1.5 effets significatifs sur 30 relèvent statistiquement de l'erreur. Ces deux effets sont de plus relativement faibles. On observe d'autre part que 8 des 20 effets significatifs obtenus impliquent la phrase utilisée. Il s'agit en conséquence d'effets contextualisés qui ne sont probablement pas généralisables.

Tableau 13: Résultas des ANOVAs à mesures répétées pour l'intensité de colère, de peur, de joie et de tristesse, effets significatifs ($p<.05$).

jugements	source	df	F	Sig.	Eta^2
intensité de peur perçue	niveau	(1, 14)	89.40	.000	0.86
	acc1	(1, 14)	6.58	.022	0.32
	acc2	(1, 14)	5.79	.030	0.29
	final	(2, 28)	20.39	.000	0.59
	phrase * niveau	(1, 14)	9.95	.007	0.42
	phrase * acc1	(1, 14)	9.80	.007	0.41
intensité de joie perçue	phrase	(1, 14)	9.04	.009	0.39
	niveau	(1, 14)	17.80	.001	0.56
	acc1	(1, 14)	7.96	.014	0.36
	acc2	(1, 14)	5.64	.032	0.29
	final	(2, 28)	13.90	.000	0.50
	phrase * acc2	(1, 14)	8.50	.011	0.38
	phrase * acc1 * acc2 * final	(2, 28)	4.18	.026	0.23
intensité de tristesse perçue	phrase * niveau * acc1	(1, 14)	5.34	.037	0.28
	acc2 * final	(2, 28)	3.92	.032	0.22
intensité de colère perçue	niveau * acc1	(1, 14)	5.38	.036	0.28
	niveau * acc2	(1, 14)	9.41	.008	0.40
	acc1 * acc2	(1, 14)	5.80	.030	0.29
	phrase * niveau * acc1 * acc2	(1, 14)	5.59	.033	0.29
	phrase * acc1 * final	(2, 28)	7.30	.003	0.34

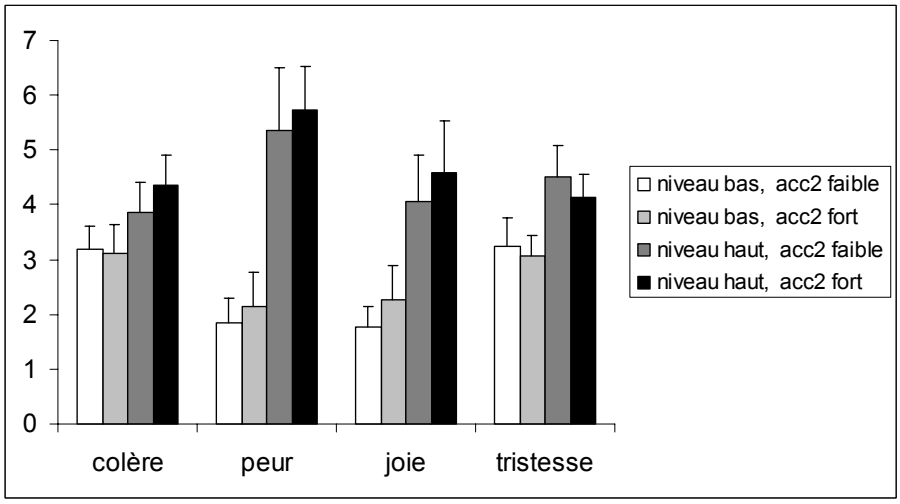
Les effets observés sont parfois difficiles à interpréter, en particulier lorsqu'il s'agit d'interactions entre 4 dimensions manipulées, et ils sont, pour la plupart, significatifs seulement dans le cadre de la variabilité restreinte par les nombreux facteurs des ANOVAs à mesures répétées. Les moyennes et les écarts-types présentés dans le graphique 13 illustrent cette dernière observation. Les ANOVAs à mesures répétées indiquent que le mouvement final de la F0 a un effet principal significatif sur l'intensité de peur et l'intensité de joie perçues. Les écarts-types représentés sur le graphique 13 correspondent à la variabilité des jugements moyens obtenus pour les 16 expressions dont le mouvement final est descendant (barres grises claires), pour les 16 expressions dont mouvement final est plat (barres grises foncées) et pour les 16 expressions dont le mouvement final est montant (barres noires). Ces écarts-types indiquent que l'intensité moyenne de joie perçue et de peur perçue pour les expressions qui présentent un même mouvement final varie très fortement. Sans restriction de la variabilité, les jugements relatifs à l'intensité de joie et de peur perçues ne seraient pas significativement affectés par le mouvement final de la F0. De plus, comme pour l'effet du niveau de base de la F0, cet effet n'est pas distinctif de l'émotion évaluée. Les expressions dont le mouvement final est montant sont perçues, en moyenne, comme exprimant une intensité de joie et de peur légèrement plus forte que les expressions dont le mouvement final est descendant.

Graphique 13: Intensité de colère, de peur, de joie et de tristesse perçues en fonction du mouvement final de la F0 (descendant, plat, montant) imposé sur les expressions de synthèse



Les contours de F0 des expressions émotionnelles examinés à la section 6.2 présentaient des variations en fonction des émotions exprimées qui semblaient refléter des différences non seulement du niveau général de la F0, mais également des différences relatives à la taille des excursions locales de F0. L'ampleur de la deuxième excursion en particulier semblait différencier les émotions exprimées. En conséquence, les effets de l'importance (faible, forte) de la deuxième excursion locale (acc2) sur les évaluations de l'intensité émotionnelle ont été plus spécifiquement examinés. Les intensités moyennes de colère, de peur, de joie et de tristesse perçues sont représentées dans le graphique 14 pour les interactions entre le niveau de base de la F0 (bas ou haut) et la deuxième excursion (acc2 faible ou fort). Les écarts-types représentent la variabilité des jugements moyens pour les 12 expressions correspondant à chaque catégorie de l'interaction.

Graphique 14: Intensité de colère, de peur, de joie et de tristesse perçues en fonction du niveau de F0 et de la taille de la deuxième excursion imposés sur les expressions de synthèse.



Le graphique 14 représente les effets principaux du niveau de F0 et de l'amplitude de la deuxième excursion sur l'intensité de peur et l'intensité de joie perçue, ainsi que l'effet d'interaction entre le niveau de F0 et l'amplitude de la deuxième excursion sur l'intensité de colère perçue. Relativement à une faible amplitude de la deuxième excursion de F0, une forte amplitude de la deuxième excursion augmente l'intensité de joie et de peur perçue. Une deuxième excursion de F0 forte n'augmente l'intensité de colère perçue que lorsque le niveau de base de la F0 est élevé.

On observe également sur le graphique 14 que la variabilité des jugements moyens pour les expressions correspondant aux différentes catégories formées par l'interaction entre le niveau de base de la F0 et l'amplitude de la deuxième excursion est assez importante. De plus, l'importance de l'effet de l'amplitude de la deuxième excursion en terme d'intensité émotionnelle perçue est relativement faible. L'intensité moyenne de colère perçue est de 3.14 pour les expressions avec un niveau de F0 bas et n'est augmentée que de 1.21 points (à 4.35 sur une échelle à 10 points) pour les expressions dont le niveau de F0 est élevé et la deuxième excursion forte. De même, la différence d'intensité émotionnelle perçue entre les expressions qui comprennent une deuxième excursion faible et les expressions qui comprennent une deuxième excursion forte n'est que de 0.50 pour l'intensité de joie perçue et de 0.34 pour l'intensité de peur perçue.

De manière plus générale, les jugements d'intensité émotionnelle obtenus pour les expressions dont les contours ont été systématiquement modifiés sont plutôt faibles (en moyenne 3.8 pour l'intensité de peur, 3.2 pour l'intensité de joie, 3.7 pour l'intensité de tristesse et 3.6 pour l'intensité de colère sur une échelle allant de 0 à 10). La faiblesse de ces jugements d'intensité ainsi qu'une fidélité inter-juges insuffisante pour une partie des jugements indiquent que les auditeurs n'ont probablement pas perçu ces expressions comme réellement émotionnelles.

6.4 Conclusions

Dans l'ensemble, il est apparu que le niveau et la variabilité de la F0 d'un ensemble d'expressions émotionnelles produites par des acteurs sont fortement affectés par les émotions exprimées, alors que des aspects plus spécifiques des contours de F0 ne varient que très peu avec les émotions exprimées. De même, il est apparu qu'il est difficile de communiquer une impression émotionnelle à des auditeurs en modifiant uniquement certains aspects des contours de F0 – telle que la hauteur de base, l'amplitude des excursions locales et le mouvement final – par la synthèse vocale. Certaines différences ont toutefois été observées. La deuxième excursion de F0 qui tend, dans nos données, à être plus proéminente pour les expressions de joie intense que pour d'autres expressions réplique notamment un résultat présenté par Pell (2001) pour des expressions émotionnelles simulées par des adultes âgés (non acteurs).

En ce qui concerne la perception de l'émotion dans la synthèse vocale, l'intensité de tristesse perçue présente un pattern de résultats qui peut sembler, à première vue, surprenant. L'intensité de tristesse perçue est en effet la seule intensité émotionnelle perçue qui n'est pas significativement dégradée par la copie croisée de la prosodie. Les expressions produites en copiant l'intonation des expressions de tristesse (2 expressions de désespoir et 2 expressions de tristesse) produites par des acteurs sur une voix de synthèse reçoivent des jugements d'intensité de tristesse qui ne sont pas beaucoup plus faibles que les jugements d'intensité de tristesse obtenus pour les expressions originales. En revanche, les manipulations systématiques des contours appliqués sur la voix de synthèse n'ont manifestement pas réussi à produire une impression de tristesse. L'intensité de tristesse évaluée pour ces expressions présente une fidélité inter-juges très faibles et les ANOVAs à mesures répétées pour tester l'effet des dimensions manipulées n'ont pas permis de mettre en évidence des influences claires des aspects des contours manipulés sur les jugements de tristesse. En résumé, les jugements de tristesse ne sont pas systématiquement influencés par les seules variations des contours de F0 alors qu'ils sont préservés par la copie de l'intonation. Cette observation pourrait s'expliquer, d'une part, par le fait que les manipulations systématiques des contours de F0 ne reproduisent pas les variations des contours qui interviennent dans la communication vocale de la tristesse effectuée par les acteurs. D'autre part, cette observation s'explique certainement aussi par la présence de l'information relative au rythme et à la durée dans les expressions produites en copiant l'intonation originale sur la voix de synthèse. Ce résultat suggèrerait donc que les aspects liés au rythme et à la durée – qui étaient constants dans les expressions systématiquement manipulées – sont particulièrement importants pour l'identification de la tristesse dans la communication vocale des émotions.

Plus généralement, les résultats obtenus indiquent que des variations simples effectuées dans les limites imposées par un système de synthèse TTS ne parviennent pas à communiquer une information émotionnelle univoque lorsqu'elles sont appliquées à un contour de F0 stylisé. Plusieurs explications peuvent être avancées à ce sujet:

Premièrement, il est possible que les variations des contours appliquées sur la voix de synthèse ne soient pas suffisamment importantes. L'étendue des variations (niveau de base élevé de 4 tons, excursions des accents augmentées de 3 tons) a été définie dans des limites qui n'introduisent pas de distorsion de la voix de synthèse. Les variations de F0 observées pour les expressions émotionnelles produites par des acteurs présentent des écarts plus importants associés à des modifications de la qualité vocale. De plus, la différence entre le niveau bas de F0 et le niveau haut est probablement plus saillante au niveau perceptif que la différence entre une excursion faible et une excursion forte;

ce qui pourrait, en partie, expliquer l'importance accordée au niveau de base de la F0 pour l'évaluation de l'intensité de peur et de joie.

D'autre part, il est probable que des variations à court terme du contour de F0 – absentes dans les expressions de synthèse systématiquement manipulées dans notre étude – jouent un rôle important dans la communication vocale des émotions. L'expérience pionnière de Lieberman & Michaels (1962) indiquait déjà que la suppression des variations fines des contours de F0 conduit à une chute considérable de la reconnaissance des émotions exprimées (les résultats rapportés par ces auteurs sont décrits dans l'introduction, section 1.4).

De plus, il reste possible que les variations de la F0 parviennent à communiquer des émotions en interaction avec d'autres caractéristiques vocales. Différents aspects de la qualité vocale, de l'intensité et des aspects rythmiques jouent probablement un rôle direct et indépendant dans la communication vocale des émotions (v. Ladd et al., 1985) mais pourraient également intervenir en interaction avec des modifications de la F0 pour produire des impressions émotionnelles.

A ce propos, Scherer, Ladd et Silverman (1984, v. aussi section 1.4) ont rapporté qu'il existe dans la littérature sur la communication non-verbale des émotions deux conceptions différentes concernant la manière dont les émotions affecteraient les expressions vocales. D'une part, les caractéristiques vocales peuvent être conçues comme covariant avec les états émotionnels exprimés (modèle de covariation). D'après ce modèle, des modifications graduelles de certaines caractéristiques vocales (par exemple un accroissement graduel de la F0 moyenne ou de l'étendue de F0) correspondraient à des modifications graduelles de l'état émotionnel d'une personne (c'est-à-dire à des émotions exprimées plus ou moins intenses) et, parallèlement, à des modifications graduelles des attributions émotionnelles effectuées par des auditeurs, indépendamment d'autres caractéristiques vocales. D'autre part, différentes caractéristiques vocales peuvent être conçues comme exprimant des émotions en interaction avec d'autres caractéristiques vocales et également avec certaines caractéristiques linguistiques des énoncés (modèle de configuration). Dans cette perspective, une configuration formée par l'association entre différentes caractéristiques vocales et un contexte linguistique spécifique produirait une impression émotionnelle spécifique. A l'appui de ce second modèle, Scherer et al. (1984) ont démontré que la forme finale des contours de F0 peut influencer les attributions émotionnelles en interaction avec la forme syntaxique des énoncés. Lorsqu'une montée finale est attendue dans un contexte grammatical défini (une question absolue) et qu'elle n'est pas réalisée, les auditeurs attribuent une émotion/attitude agressive au locuteur.

Les donnée présentée ci-dessus apportent un soutien incontestable au modèle de covariation. En effet, les différences principales observées – aussi bien pour de la description des contours de F0

produits par les acteurs que pour la perception des expressions manipulées – concernent la hauteur et la variabilité globales de la F0. Indépendamment d'autres caractéristiques vocales, la hauteur du niveau de base de la F0 ainsi que la hauteur relative des excursions locale varient dans les données présentées en fonction du niveau d'activation associé aux émotions exprimées. Dans le même sens, pour les expressions dont les contours de F0 ont été manipulés, la hauteur du niveau de base correspond à la caractéristique qui a influencé le plus fortement les attributions émotionnelles; les expressions avec un niveau de F0 relativement élevé ont été jugées plus émotionnelles que les expressions dont le niveau de F0 était plus bas.

Les résultats présentés ci-dessus n'ont en revanche pas permis de mettre en évidence des configurations spécifiques de la F0 pour différentes émotions exprimées. Tout au plus certaines tendances sont apparues, suggérant notamment que le mouvement final des contours et la hauteur relative de la deuxième excursion pourraient varier en fonction des émotions exprimées. Cela n'exclut pas, évidemment, que des configurations entre des contours spécifiques de la F0 et des aspects syntaxiques (tels que décrits par Scherer et al., 1984) ou encore des aspects sémantiques ou éventuellement phonétiques puissent intervenir dans la communication non-verbale des émotions. Les expressions émotionnelles produites par des acteurs que nous avons analysées ne comportent pas de contenu sémantique ou syntaxique qui permettrait de communiquer une impression émotionnelle en interaction avec des aspects de la F0. Toutefois, les excursions locales de F0 ("accents") produites sur ces expressions ont été réalisées avec une grande variabilité relativement à leur position exacte dans les énoncés. Cette variabilité a été éliminée par la stylisation de la F0. Il reste donc possible qu'un examen plus approfondi de la position des excursions relativement au contenu phonétique permette de mettre en évidence des différences spécifiques à des émotions exprimées ou perçues.

7 Synthèse et perspectives

Cette dernière section vise à présenter, dans un premier temps, une synthèse des études effectuées et des principaux résultats obtenus. Plusieurs problèmes soulevés par ces études seront ensuite développés. Finalement, quelques propositions seront formulées relativement aux principales perspectives de recherche dans ce domaine.

Un nombre relativement important de problèmes et de résultats spécifiques, analysés et discutés dans les sections précédentes, ne seront pas repris dans cette section qui vise avant tout à présenter une *synthèse* des résultats et des difficultés rencontrées dans cette recherche. Les problématiques, les résultats et les conclusions ponctuel(le)s – relatif(ve)s à des aspects spécifiques aux différentes sections de cette thèse – ont été préalablement examiné(e)s dans les discussions et les conclusions présentées à l'issue des sections précédentes.

7.1 Synthèse de la recherche effectuée et des principaux résultats obtenus

La structure de cette synthèse reproduit la structure de la recherche présentée dans la thèse. Les approches utilisées et les principaux résultats obtenus sont résumés d'abord pour les caractéristiques acoustiques des expressions émotionnelles, puis pour les caractéristiques vocales perçues des expressions et pour les attributions émotionnelles effectuées relativement aux expressions vocales. Les principaux résultats obtenus pour la modélisation de l'ensemble du processus de communication sont présentés ensuite. Finalement, les approches et les principaux résultats obtenus relativement à la participation de l'intonation dans la communication vocale des émotions seront rappelés.

7.1.1 Caractéristiques acoustiques des expressions émotionnelles (section 2)

Un grand nombre de paramètres acoustiques ont été extraits de 144 expressions vocales produites par 9 acteurs qui ont exprimé (simulé) 8 émotions (colère chaude et froide, anxiété et peur panique, joie calme et exaltation, tristesse et désespoir) en prononçant 2 énoncés sans signification. Des paramètres acoustiques relatifs à la fréquence fondamentale, à l'intensité, à la durée et à la distribution spectrale de l'énergie ont été mesurés. Afin de contrôler les variations imputables aux locuteurs, les paramètres acoustiques extraits ont été standardisés pour chaque locuteur. Des analyses de variance ont permis d'évaluer l'influence des émotions et des énoncés sur ces paramètres standardisés. La présence d'une forte colinéarité entre les paramètres acoustiques a conduit à sélectionner un nombre réduit de paramètres sur la base d'une analyse en composantes principales. La possibilité de discriminer les émotions exprimées sur la base de 8 paramètres acoustiques sélectionnés (l'intensité moyenne, l'étendue d'intensité, la F0 moyenne, le minimum de

F0, la durée totale, la durée relative des segments voisés, la proportion d'énergie inférieure à 100Hz et la proportion d'énergie comprise entre 600 et 800 Hz) a été évaluée au moyen d'une analyse discriminante. Finalement, l'influence respective du type d'émotion exprimé et de l'activation sous-jacente aux émotions exprimées (définie comme faible pour les expressions de colère froide, d'anxiété, de joie calme et de tristesse et comme forte pour les expressions de colère chaude, de peur panique, d'exaltation et de désespoir) a été examinée en contrôlant statistiquement l'influence du niveau d'activation sur les paramètres acoustiques.

Les résultats obtenus indiquent que la majorité des paramètres acoustiques mesurés sont fortement affectés par les 8 émotions exprimées. Le contrôle du niveau d'activation sous-jacent aux émotions atténue sensiblement l'effet des émotions sur les paramètres acoustiques; le type d'émotion exprimé (colère, peur, joie ou tristesse) conserve toutefois une influence sur une partie des paramètres acoustiques indépendamment du niveau d'activation dichotomique (faible ou fort). L'analyse discriminante (cross validated) effectuée sur la base des 8 paramètres acoustiques sélectionnés a permis de reclasser correctement 50% des expressions vocales dans les 8 catégories émotionnelles.

Dans l'ensemble, les résultats suggèrent que les paramètres acoustiques mesurés reflètent avant tout le niveau d'activation sous-jacent aux expressions émotionnelles, mais sont également capables, dans une moindre mesure, de différencier différents types (familles) d'émotions exprimées.

7.1.2 Caractéristiques vocales perçues des expressions émotionnelles (section 3)

Plusieurs prétests consacrés à évaluer la possibilité d'obtenir des jugements relatifs à différentes caractéristiques vocales perçues ont abouti à la sélection de 8 dimensions vocales correspondant à: la hauteur, l'intensité, la rapidité, l'intonation, l'instabilité, la qualité de l'articulation, la qualité rauque et perçante. Une procédure destinée à améliorer la fidélité des jugements relatifs aux caractéristiques vocales perçues a été développée. Dans cette procédure, les jugements relatifs aux dimensions vocales sélectionnées sont effectués en comparant directement les expressions émotionnelles produites par chaque locuteur sur chaque dimension vocale. Plusieurs groupes d'auditeurs/juges ont été recrutés afin d'évaluer les 144 expressions émotionnelles examinées sur les 8 dimensions vocales. La fidélité inter-auditeurs a été ensuite évaluée pour les jugements obtenus et des jugements moyens ont été calculés pour les expressions émotionnelles sur les 8 dimensions vocales. Les relations entre les jugements moyens relatifs aux 8 dimensions vocales et un ensemble de paramètres acoustiques ont été évaluées. Des analyses de variance ont été effectuées afin de tester l'influence des émotions exprimées sur les caractéristiques vocales perçues. La possibilité de discriminer les 8 émotions exprimées à l'aide des 8 caractéristiques vocales perçues a été évaluée par une analyse discriminante. Finalement, l'influence respective du niveau d'activation et du type

d'émotion exprimée sur les caractéristiques vocales perçues a été évaluée en contrôlant statistiquement l'influence du niveau d'activation sur les jugements relatifs aux caractéristiques vocales perçues.

La fidélité inter-auditeurs des jugements varie considérablement en fonction des dimensions vocales considérées. L'accord inter-auditeurs est très élevé pour certaines dimensions (e.g. l'intensité perçue) et beaucoup plus faible pour d'autres dimensions (e.g. la qualité de l'articulation). Dans l'ensemble, les coefficients de fidélité sont toutefois satisfaisants. Les paramètres acoustiques sélectionnés sont relativement fortement corrélés avec une partie des caractéristiques vocales perçues (l'intensité, la rapidité, la hauteur, l'intonation et la qualité perçante). En revanche, ces paramètres acoustiques ne parviennent pas à rendre compte, dans la même mesure, des jugements d'instabilité et des jugements relatifs à la qualité rauque et à la qualité de l'articulation. Les jugements moyens obtenus pour les 8 dimensions vocales évaluées sont assez fortement affectés par l'émotion exprimée. L'analyse discriminante effectuée sur la base des 8 caractéristiques vocales perçues a permis de reclasser correctement 61% des expressions émotionnelles dans les 8 catégories émotionnelles d'origine. Le contrôle statistique du niveau d'activation sous-jacent aux émotions exprimées diminue l'influence de l'émotion exprimée sur les caractéristiques vocales perçues, mais ne la supprime pas. Les 5 caractéristiques vocales perçues fortement corrélées avec les paramètres acoustiques sont plus fortement influencées par le niveau d'activation que les 3 caractéristiques vocales perçues moins fortement corrélées aux paramètres acoustiques.

La discrimination des émotions exprimées par les caractéristiques vocales perçues s'est avérée plus performante que la discrimination réalisée par les paramètres acoustiques dans la première section. Relativement à ce résultat, deux explications ont été proposées: (1) Certaines caractéristiques vocales perçues permettraient de représenter des caractéristiques vocales importantes pour la discrimination des émotions exprimées et qui ne seraient pas représentées sur le plan des mesures acoustiques effectuées. (2) La reconnaissance des émotion exprimée dominerait la perception des expressions vocale et orienterait (biaiserait) les jugements relatifs aux caractéristiques vocales perçues. Les données disponibles ne permettent pas d'exclure une telle influence des émotions perçues sur les évaluations relatives aux caractéristiques vocales perçues.

7.1.3 Attributions émotionnelles (section 4)

Des jugements relatifs à l'intensité perçue de tristesse, de joie, de peur et de colère ont été recueillis pour les 144 expressions émotionnelles. Pour chaque intensité émotionnelle perçue, les jugements ont été obtenus en comparant directement les expressions produites par un même locuteur/acteur. La fidélité inter-auditeurs a été évaluée pour chaque jugement d'intensité émotionnelle et des

jugements moyens ont été calculés pour chaque expression émotionnelle. L'influence globale des émotions exprimées (8 niveaux, within) ainsi que l'influence des différents locuteurs/acteurs (9 niveaux, between) sur les jugements d'intensité émotionnelle ont été évaluées par des analyses de variance. Les influences respectives du type d'émotion exprimée (joie, peur, colère ou tristesse) et du niveau d'activation (fort ou faible) ont également été testées par des analyses de variance.

La fidélité inter-auditeurs pour les quatre jugements d'intensité émotionnelle a été jugée satisfaisante. Les résultats des analyses de variance indiquent que l'effet des émotions exprimées sur les quatre jugements moyens d'intensité émotionnelle est très important. Dans l'ensemble, les effets directs des locuteurs sont faibles, alors que les effets d'interaction entre le locuteur et l'émotion exprimée sont importants. Certaines tendances observées suggèrent que certains locuteurs parviennent à communiquer un type d'émotion alors qu'ils échouent à communiquer un autre type d'émotion.

Le niveau d'activation et le type d'émotion exprimée affectent de manière similaire les jugements d'intensité de peur, de joie et de colère: l'intensité émotionnelle perçue est plus faible pour les expressions faiblement activée (anxiété, joie calme et colère froide) et plus forte pour les expressions fortement activée (peur panique, joie intense et colère chaude). En revanche, les jugements d'intensité de tristesse ne sont significativement affectés que par le type d'émotion exprimée et non par le niveau d'activation; les expressions de 'tristesse désespérée' ne sont pas perçues comme exprimant une intensité de tristesse plus importante que les expressions de 'tristesse déprimée'. Une forme de confusion est apparue pour les expressions de désespoir qui sont perçues non seulement comme exprimant une intensité de tristesse élevée, mais également comme exprimant une intensité de peur élevée. De manière générale, les émotions exprimées sont toutefois relativement bien "reconnues".

7.1.4 Modélisation du processus de communication (section 5)

Deux modèles statistiques (Lens Model Equation et Path Analysis) ont été appliqués à l'analyse du processus de communication vocale des émotions.

La Lens Model Equation (LME) décompose la relation (corrélation) entre l'émotion exprimée et l'émotion perçue en deux composantes: une composante modélisée par les caractéristiques vocales (acoustiques ou perçues) et une composante non-modélisée. La composante modélisée est elle-même composée de trois composantes: la relation entre les caractéristiques vocales et l'émotion exprimée (corrélation multiple), la relation entre les caractéristiques vocales et l'émotion perçue (corrélation multiple), et la relation entre l'utilisation des caractéristiques vocales dans la

modélisation de l'émotion exprimées et l'utilisation des caractéristiques vocales dans la modélisation de l'émotion perçue (corrélation des valeurs prédites par les régressions multiples).

La Path Analysis décompose la relation entre l'émotion exprimée et l'émotion perçue en trois composantes: les effets centraux (médiatisés par les paramètres acoustiques et les caractéristiques vocales perçues), les effets périphériques (médiatisés uniquement par les paramètres acoustiques ou médiatisées uniquement par les caractéristiques vocales perçues), l'effet direct (part de la relation qui n'est pas médiatisée par les caractéristiques vocales qui figurent dans le modèle).

Dans ces deux modèles, l'émotion exprimée a été codée de manière dichotomique (absence/présence). Cette définition de l'émotion exprimée introduit un contrôle de l'effet du niveau d'activation (pour chaque modalité de la variable dichotomique, une moitié des expressions sont fortement activées et l'autre moitié faiblement activées).

Huit modélisations basées sur la LME ont été réalisées dans un premier temps: les relations entre la colère exprimée et perçue, entre la joie exprimée et perçue, entre la peur exprimée et perçue, et entre la tristesse exprimée et perçue ont été modélisée sur la base des 8 paramètres acoustiques sélectionnés dans la section 2, puis sur la base des 8 caractéristiques vocales perçues évaluées dans la section 3. Dans un deuxième temps, deux modélisations de la relation entre le niveau d'activation exprimé (variable dichotomique) et la moyenne de l'intensité émotionnelle perçue (un 1er modèle basé sur les 8 paramètres acoustiques et un 2ème modèle basé sur les 8 caractéristiques vocales perçues) ont été ajoutées aux modèles réalisés pour la communication des 4 catégories émotionnelles.

Cinq modèles basés sur la Path Analysis ont été réalisés. Pour chaque modèle, trois paramètres acoustiques et trois caractéristiques vocales perçues ont été sélectionnées sur la base de leurs relations avec les émotions (ou l'activation/intensité) exprimées et perçues. Trois paramètres acoustiques et trois caractéristiques vocales perçues ont donc été utilisé(e)s conjointement pour la médiation de la relation entre la colère exprimée et la colère perçue, entre la joie exprimée et perçue, entre la peur exprimée et perçue, entre la tristesse exprimée et perçue, et entre le niveau d'activation exprimé et l'intensité émotionnelle moyenne perçue.

Les résultats de l'analyse par la LME ont montré que la composante modélisée de la communication des quatre émotions (colère, peur, joie et tristesse) est plus importante dans les modèles basés sur les caractéristiques vocales perçues que dans les modèles basés sur les paramètres acoustiques. Les relations entre les caractéristiques vocales (acoustiques et perçues) et les émotions exprimées sont très légèrement plus faibles que les relations entre les caractéristiques vocales (acoustiques et perçues) et les émotions perçues. Dans la plupart des modèles, l'utilisation des caractéristiques

vocales à l'encodage et l'utilisation des caractéristiques vocales au décodage sont très similaires. La capacité des caractéristiques vocales (les paramètres acoustiques ou les caractéristiques vocales perçues utilisées dans les modèles) à rendre compte de la communication varie en fonction des émotions communiquées. Les paramètres acoustiques, par exemple, ne parviennent pas à rendre compte de la communication de la joie (3% de la relation entre la joie exprimée et la joie perçue a pu être modélisée), mais parviennent à rendre compte en partie de la communication de la colère (58% de la relation entre la colère exprimée et la colère perçue a pu être modélisée). Les caractéristiques vocales utilisées parviennent d'autre part à rendre compte de la quasi-totalité de la relation entre l'activation exprimée et l'intensité émotionnelle moyenne perçue (85% de la relation modélisée par les paramètres acoustiques et 94% de la relation modélisée par les caractéristiques vocales perçues).

Pour la Path Analysis, les poids respectifs des différents effets varient en fonction de l'émotion communiquée. Les effets centraux ne sont relativement importants que pour la communication de la colère (ils représentent environ un tiers de la relation totale entre la colère exprimée et la colère perçue). Pour les quatre émotions communiquées, les effets périphériques sont davantage basés sur les caractéristiques vocales perçues que sur les paramètres acoustiques et les effets les plus importants sont les effets directs qui représentent environ deux tiers de la relation totale pour la colère, la peur et la tristesse et environ 85% de la relation totale entre la joie exprimée et la joie perçue. Les effets directs très importants dans ces modèles signalent que les caractéristiques vocales sélectionnées (trois paramètres acoustiques et trois caractéristiques vocales perçues) ne parviennent pas à médiatiser la plus grande partie de la relation entre les émotions exprimées et les émotions perçues. En revanche, les caractéristiques vocales sélectionnées pour médiatiser la relation entre l'activation exprimée et l'intensité émotionnelle moyenne perçue parviennent à médiatiser la quasi-totalité de cette relation.

Le fait que les caractéristiques vocales (acoustiques et perçues) ne parviennent à rendre compte que d'une partie relativement faible de la communication émotionnelle – en particulier pour la communication de la joie mais également pour les trois autres types d'émotions – suggère que les caractéristiques vocales (acoustiques et perçues) mesurées ne sont probablement pas appropriées pour rendre compte de la communication émotionnelle. L'hypothèse que l'évaluation d'autres caractéristiques vocales, plus appropriées, permettraient de rendre compte d'une part plus importante de la communication émotionnelle a été formulée.

7.1.5 La contribution de l'intonation (section 6)

Une stylisation manuelle du contour de F0 pour les 144 expressions émotionnelles examinées a été effectuée en relevant un ensemble de points clés des contours de F0. Différentes mesures ont été dérivées de ces points (e.g. pentes pour des montées et des descentes locales de F0, hauteurs relatives de différents pics de F0). Les effets des émotions exprimées sur les mesures dérivées des contours de F0 ont été testés en effectuant des analyses de variance. La possibilité de rendre compte des différences observées pour les émotions exprimées en faisant appel à des mesures plus globales de la F0 (telles que la F0 moyenne) a été examinée en contrôlant statistiquement les effets de ces mesures globales sur les mesures dérivées de contours de F0.

Afin d'évaluer l'influence des variations du contour de F0 sur les attributions émotionnelles, des expressions de synthèse (text-to-speech) ont été produites avec des manipulations systématiques de plusieurs aspects des contours de F0. Les aspects manipulés incluent la hauteur des contours (deux niveaux, bas/haut), la hauteur de deux excursions locales (deux niveaux, faible/fort), la forme du mouvement final (3 niveaux, montant/descendant/plat). Ces caractéristiques ont été manipulées pour deux énoncés sans signification et réalisées avec deux voix de synthèse. Au total, 96 expressions de synthèse ont été produites. Des études de jugements ont été effectuées dans le but d'évaluer l'influence de ces manipulations sur les attributions émotionnelles. La procédure d'évaluation utilisée dans la section 3 et la section 4 a été utilisée pour recueillir des jugements relatifs à l'intensité de peur, de tristesse, de colère et de joie perçue dans les expressions de synthèse. D'autre part, 16 expressions produites par les acteurs ont été sélectionnées et des expressions produites par une copie croisée de l'intonation (entre l'intonation du système de synthèse et l'intonation des expressions produites par les acteurs) ont été produites. Des attributions émotionnelles ont été obtenues pour les 32 expressions résultantes en utilisant la même procédure que pour les expressions systématiquement manipulées.

Les résultats obtenus pour la stylisation des contours de F0 produits par les locuteurs/acteurs indiquent que les propriétés des contours examinées sont affectées principalement par le niveau d'activation sous-jacent aux émotions exprimées. Le niveau de hauteur global des contours et la taille relative des excursions locales varient avec le niveau d'activation. Pour la majorité des mesures extraites des contours de F0, les différences observées entre différentes émotions exprimées – ou, le plus souvent, entre différents niveaux d'activation – peuvent être réduites à des variations plus globales des contours (telles que la F0 moyenne ou l'étendue de F0).

Les études de jugements indiquent que les expressions de synthèse (text-to-speech) dont les contours ont été systématiquement modifiés ne sont pas perçues comme étant très émotionnelles.

Quelques aspects manipulés affectent néanmoins significativement les jugements d'intensité émotionnelle. Les expressions dont le contour est globalement plus élevé sont notamment perçues comme communicant une intensité émotionnelle plus forte (en particulier pour la peur et la joie). Les résultats pour les expressions produites par la copie croisée de l'intonation indiquent que les expressions des locuteurs/acteurs resynthétisées avec le contour d'intonation "neutre" du système de synthèse sont perçues comme plus émotionnelles que les expressions de synthèse (text-to-speech) sur lesquelles les contours d'intonation des expressions produites par les locuteurs/acteurs ont été copiés. De plus, la qualité émotionnelle des expressions vocales (colère, peur, joie ou tristesse) est mieux préservée dans les expressions de resynthèse (qui conservent la qualité vocale des expressions originales) que dans les expressions de synthèse qui reproduisent les contours d'intonation produits par les acteurs (avec la qualité vocale du système de synthèse).

La forme des contours de F0 ne semble donc pas contribuer de manière très importante à la communication des émotions étudiées. Il reste toutefois possible que certains aspects des contours qui n'ont pas été examinés – par exemple l'interaction des excursions de F0 avec le contenu phonétique ou, plus généralement, l'interaction des contours de F0 avec le contenu linguistique des énoncés – jouent un rôle important sur le plan de la communication vocale des émotions.

7.2 Aspects critiques de la recherche présentée

Différents types de problèmes ont été rencontrés au cours des études effectuées. Un certain nombre de critiques ont notamment été formulées dans les discussions et les conclusions qui clôturent chaque section de la recherche présentée ci-dessus. Quelques problèmes et quelques critiques d'ordre général seront développés ci-dessous. Nous avons choisi de présenter successivement des problématiques associées (1) aux expressions vocales utilisées et à la définition des émotions exprimées; (2) aux mesures des caractéristiques vocales (acoustiques et perçues) effectuées; (3) à la présentation des résultats et aux méthodes statistiques choisies. Ces différentes problématiques sont présentées séparément et successivement dans un but de structuration; elles ne sont cependant pas conçues comme étant indépendantes.

7.2.1 La nature des expressions vocales étudiées

Nous avons choisi d'étudier des expressions vocales correspondant à 4 types (familles) d'émotions – colère, peur, joie et tristesse – et présentant deux niveaux d'activation (faible versus fort). Au cours des études effectuées au moins deux problèmes relatifs à la définition des émotions étudiées sont apparus.

Premièrement, la *définition dichotomique de l'activation* pourrait être insuffisante, au sens où des niveaux d'activation différents sont probablement recouverts par le niveau défini comme faible, ainsi que par le niveau défini comme fort. Nous pensons en particulier que le niveau d'activation sous-jacent aux expressions de tristesse que nous avons étudiées est en général plus faible que le niveau d'activation sous-jacent aux expressions de colère froide que nous avons examinées; ceci bien que dans notre opérationnalisation les expressions de colère froide et de tristesse soient définies comme correspondant à un même niveau d'activation 'faible' (par opposition aux expressions de colère chaude et de désespoir, conçues comme fortement activées).

Plus globalement, l'opérationnalisation dichotomique du niveau d'activation – réalisée sur la base de postulats théoriques relativement à la présence d'une activation forte ou faible – ne représente pas la manière idéale d'étudier l'influence du niveau d'activation émotionnelle sur les expressions vocales. L'effet de l'activation – conçue comme une *dimension continue* sous-jacente aux réactions émotionnelles – sur les expressions vocales devrait être idéalement examiné dans un contexte où une mesure indépendante (et continue) du niveau d'activation serait disponible. Deux approches différentes pourraient permettre d'obtenir une telle mesure: d'une part, une procédure dans laquelle les locuteurs indiqueraient le niveau d'activation (auto-évalué) associé aux émotions qu'ils expriment pourrait être envisagée; d'autre part, des mesures physiologiques (e.g. conductance cutanée, température, rythme cardiaque) pourraient être utilisées de manière à évaluer le niveau d'activation physiologique associé à différentes expressions vocales.

Dans une définition conservatrice, il conviendrait donc de ne pas appliquer la notion de *niveau d'activation* aux différentes expressions étudiées dans cette thèse, mais de considérer, plus modestement, que des expressions correspondant à deux sortes d'émotions (e.g. la colère froide et la colère chaude) habituellement regroupés dans une seule catégorie "fondamentale" (e.g. la colère) ont été systématiquement examinées pour 4 catégories "fondamentales". La distinction effectuée à l'intérieur de ces "catégories fondamentales" reste néanmoins définie sur le plan d'une activation émotionnelle théoriquement plus ou moins importante.

Un second problème de définition est apparu spécifiquement pour les expressions de 'désespoir' qui ont été définies théoriquement comme correspondant à une forme de 'tristesse' fortement activée. Dans ce cas, le problème est lié à la *catégorie/famille* émotionnelle à laquelle les expressions de 'désespoir' appartiennent effectivement. Les scénarios utilisés pour définir les expressions de 'désespoir' comportent une composante d'urgence qui rapproche ces situations de situations qui pourraient éventuellement évoquer une association de peur et de tristesse (v. annexe A2). De plus, les expressions de 'désespoir' sont jugées par les auditeurs ayant participé à cette recherche comme

exprimant une intensité de peur et une intensité de tristesse relativement élevées (v. section 4). Les expressions de 'désespoir' examinées *communiquent* donc simultanément une part de peur et une part de tristesse. Du fait de la confusion (entre peur et tristesse) introduite par les scénarios définissant le 'désespoir', les données disponibles ne nous permettent pas d'évaluer dans quelle mesure les acteurs auraient été capables de produire des expressions avec une activation forte qui auraient été perçues comme exprimant uniquement de la tristesse (ou si les expressions tristes fortement activées sont dans tous les cas perçues comme exprimant également de la peur).

D'autre part, les caractéristiques vocales associées aux expressions de 'désespoir' sont plus proches de caractéristiques vocales qui ont été rapportées ailleurs comme reflétant de la 'peur', que de propriétés vocales habituellement considérées comme caractéristiques de la 'tristesse'. Dans la mesure où les expressions de 'désespoir' sont perçues non seulement comme exprimant de la peur mais également comme exprimant de la tristesse, cette observation ne permet cependant pas de conclure que les expressions de 'désespoir' exprimeraient uniquement de la peur et non de la tristesse. Les résultats indiquent au contraire que la tristesse peut être communiquée dans des expressions de tristesse non-prototypique (la tristesse prototypique est faiblement activée) qui communiquent également de la peur.

Plus globalement, cette "confusion" entre tristesse et peur dans les expressions de 'désespoir' soulève la question de la pertinence de la conceptualisation des émotions proposée dans la recherche présentée. La définition des émotions proposée – qui croise 4 catégories "fondamentales" avec deux niveaux d'activation – présente l'avantage de distinguer deux sortes d'émotions (fortement versus faiblement activées) à l'intérieur des catégories "fondamentales" habituellement considérées. L'utilisation des catégories "fondamentales" conduit cependant dans le cas présent à invoquer la présence éventuelle d'un "mélange" entre deux catégories "fondamentales". Or la pertinence du concept de "mélange" entre états émotionnels sur le plan des expressions vocales est questionnable. S'il est relativement aisé de concevoir un mélange entre deux expressions faciales – qui comprendrait des traits caractéristiques associés habituellement à chacune des deux expressions – il est en revanche beaucoup plus difficile d'imaginer comment se traduirait un tel mélange sur le plan vocal. A notre connaissance, il n'existe aucune proposition concrète dans la littérature à ce propos. Cette observation souligne une nouvelle fois la nécessité de développer une taxonomie (ou un système de description dimensionnel) des émotions adaptée au domaine de l'étude des expressions vocales émotionnelles (et qui permettrait de faire l'économie du concept de "mélange" entre "émotions fondamentales").

D'autres problèmes pourraient encore être évoqués relativement aux expressions vocales utilisées dans cette recherche. Les problèmes associés notamment à l'utilisation d'expressions simulées par des acteurs (qui produisent des expressions très prototypiques, peu nuancées) ou encore les problèmes associés à l'absence de contexte verbal et extra-verbal sont évidemment présents dans les études que nous avons réalisées. Ces problèmes sont toutefois communs à la majorité des études effectuées dans ce domaine et ont été longuement discutés dans d'autres travaux. Nous nous contenterons de rappeler ici que dans cette recherche, tout comme dans les autres travaux effectués dans ce domaine, les résultats sont évidemment totalement dépendants des expressions vocales sélectionnées. La généralisation des résultats ne doit jamais être effectuée légèrement, en particulier lorsqu'un certain nombre d'aspects des conditions expérimentales sont peu représentatives des conditions habituelles de communication (e.g. la simulation par des acteurs, l'absence de contexte verbal et extra-verbal).

7.2.2 Les mesures des caractéristiques vocales effectuées

Dans la recherche présentée, des paramètres acoustiques ont été mesurés (section 3) et des caractéristiques vocales perçues ont été évaluées (section 4). Pour ces deux types de mesures, les aspects de la voix qui ont été évalués ont été en partie définis par différentes contraintes liées à des aspects pratiques ou à la structure des données obtenues.

En ce qui concerne les paramètres acoustiques, la sélection originale des 44 paramètres mesurés a été définie relativement aux paramètres qui sont habituellement mesurés dans le domaine de l'étude des expressions vocales émotionnelles. Plus spécifiquement, les paramètres que nous avons extraits correspondent aux paramètres mesurés dans l'étude de Banse et Scherer (1996) pour des expressions vocales provenant de la même base de données que les expressions que nous avons examinées. La forte colinéarité entre une partie de ces paramètres nous a conduit à sélectionner un nombre réduit de paramètres. Cette sélection a été effectuée de manière empirique en identifiant des paramètres relativement indépendants par le biais d'une analyse en composantes principales. Nous avons néanmoins inclus un paramètre supplémentaire qui n'a pas été sélectionné dans le cadre de l'identification des composantes sous-jacentes aux expressions vocales mesurées: l'intensité moyenne a été ajoutée aux paramètres sélectionnés empiriquement, sur la base de son importance théorique dans la communication vocale des émotions.

En ce qui concerne les caractéristiques vocales perçues, la sélection des aspects vocaux évalués a été limitée par la nécessité de s'assurer de la bonne compréhension des dimensions vocales évaluées par des auditeurs/juges. Des prétests ayant démontré qu'un grand nombre de dimensions vocales distinguées dans la littérature (par exemple creaky/grinçante ou breathy/soufflée) ne sont pas

comprises par des auditeurs non-experts de la qualité vocale, seules 8 caractéristiques vocales dont la définition semblaient être relativement bien partagée ont été soumises à l'évaluation des auditeurs/juges. Certains aspects vocaux qui semblaient pouvoir être identifiés fiablement – en particulier la qualité nasale/dénasalisée de la voix ou le dévoisement (chuchotement) – n'ont pas été évaluées dans la mesure où nous avons jugé que ces aspects ne variaient pas suffisamment dans les expressions vocales examinées (en d'autres termes, nous avons estimé que ces aspects n'étaient probablement pas caractéristiques de différentes expressions émotionnelles).

La sélection des mesures (acoustiques et perçues) a donc été effectuée en partie sur la base de postulats théoriques relativement aux caractéristiques vocales susceptibles d'intervenir (ou non) sur le plan de la communication vocale des émotions, mais également en partie sur la base de contraintes méthodologiques ou relativement à la structure empirique des données (en ce qui concerne la sélection des paramètres acoustiques). L'absence d'un modèle théorique global relativement aux caractéristiques acoustiques et aux caractéristiques vocales perçues intervenant dans la communication vocale des émotions constitue probablement le défaut le plus important de la recherche effectuée. Un modèle théorique plus cohérent relativement aux caractéristiques vocales (acoustiques et perçue) impliquées dans la communication vocale des émotions aurait notamment pu nous conduire à mesurer des aspects acoustiques et des aspects vocaux perçus correspondant plus systématiquement à des caractéristiques vocales conceptuellement équivalentes.

L'absence de modèle relatif aux caractéristiques vocales impliquées dans la communication émotionnelle peut s'expliquer à plusieurs niveaux. D'une part, un tel modèle n'a, à ce jour, pas encore été formulé. A notre connaissance, il n'existe qu'un seul modèle théorique relatif aux caractéristiques vocales qui seraient associées aux réactions émotionnelles. Il s'agit du modèle proposé par Scherer (1986, 2003) qui relie les critères d'évaluation cognitive à l'origine des réactions émotionnelles proposés par cet auteur à des réactions physiologiques, elles-mêmes associées à des modifications des caractéristiques vocales. Ce modèle est donc centré sur les processus de production des expressions vocales associés à la composante physiologique des réactions émotionnelles. Dans le contexte de la recherche présentée, l'accent a été mis davantage sur le processus de communication et les états émotionnels communiqués ont été définis dans le cadre d'une conceptualisation qui fait appel à la notion d'émotions "fondamentales" et à la notion d'activation émotionnelle (et non à des dimensions d'évaluation cognitive ou à des patterns de réponses physiologiques plus spécifiques). Le modèle proposé par Scherer ne peut être aisément transposé à ce contexte.

Les autres approches proposées dans le domaine de l'étude des expressions vocales émotionnelles sont essentiellement exploratoires et théoriquement peu explicites. La plupart des études sont réalisées afin de tester la possibilité de reconnaître des émotions exprimées ou afin d'établir empiriquement des profils acoustiques pour différents types d'émotions exprimées, elles ne visent pratiquement jamais à tester des prédictions relativement aux caractéristiques vocales impliquées dans la communication émotionnelle.

Par ailleurs, les contraintes/limitations pratiques sur les caractéristiques vocales (acoustiques ou perçues) qu'il est possible de mesurer/évaluer sont bien réelles. Dans le cas où un modèle permettrait de définir les caractéristiques vocales théoriquement importantes pour la communication vocale des émotions, il ne serait pas nécessairement techniquement (pratiquement) possible de mesurer/évaluer les caractéristiques vocales que l'on voudrait théoriquement examiner. Dans le cadre de la recherche présentée, nous étions par exemple intéressés à obtenir des évaluations de la qualité soufflée (aspirée) des expressions vocales qui est à notre avis une caractéristique importante dans le domaine de la communication vocale des émotions. Cette caractéristique est cependant difficile à évaluer à la fois sur le plan de ses corrélats acoustiques et sur le plan des jugements perceptifs.

Indépendamment de ces limitations pratiques, le domaine de l'étude des expressions vocales émotionnelles bénéficierait toutefois indiscutablement du développement de propositions théoriques et du développement de mesures (acoustiques et perçues) plus spécifiquement adaptées à ce domaine.

7.2.3 La présentation des résultats et les méthodes statistiques employées

Un nombre relativement important de méthodes statistiques ont été mises en oeuvre pour analyser les résultats présentés dans cette recherche. L'objectif de cette section n'est pas d'examiner et encore moins de mettre en cause systématiquement toutes les méthodes employées, mais plutôt de considérer certains aspects qui émergent de la comparaison de différentes méthodes statistiques utilisées dans différentes sections de la thèse. Dans un deuxième temps, quelques considérations d'ordre général relativement à l'interprétation des résultats obtenus par le biais de l'approche corrélationnelle seront formulées.

Dans les sections 2 et 3, des analyses de variance ont été utilisées afin d'évaluer l'effet des émotions exprimées (8 niveaux) sur les 8 paramètres acoustiques sélectionnés et sur les 8 caractéristiques vocales perçues qui ont été évaluées. Des analyses discriminantes ont également été effectuées dans ces sections afin d'évaluer la possibilité de discriminer les 8 catégories émotionnelles à l'aide des paramètres acoustiques et des caractéristiques vocales perçues.

Les analyses de variance ont été effectuées de manière à tester uniquement l'effet de l'émotion exprimée sur les paramètres acoustiques et les caractéristiques vocales perçue, en contrôlant les effets qui pourraient être attribués aux énoncés prononcés ou aux locuteurs, ainsi que les effets d'interaction entre les énoncés et les émotions exprimées et les effets d'interaction entre les émotions exprimées et les locuteurs. Ces analyses ont mis en évidence des effets très importants des émotions exprimées sur les paramètres acoustiques et sur les caractéristiques vocales perçues.

Les études effectuées dans ce domaine se contentent habituellement de rapporter les moyennes (et les écarts-types) de chaque caractéristique vocale examinée (en général des paramètres acoustiques) pour chaque type d'émotion considéré. Les analyses discriminantes ont été effectuées avec comme objectif essentiel d'ajouter à ce type de description une évaluation du pouvoir de discrimination conjugué de l'ensemble des paramètres acoustiques et de l'ensemble des caractéristiques vocales perçues relativement aux émotions exprimées. Ces analyses discriminantes, dans lesquelles la variabilité introduite par les différents locuteurs[20] et les différents énoncés n'est pas contrôlée, ont indiqué que les paramètres acoustiques d'une part et les caractéristiques vocales perçues d'autre part parviennent à réaliser une assez bonne discrimination de la plupart des catégories émotionnelles. Certaines catégories émotionnelles sont toutefois moins bien discriminées que d'autres. Les paramètres acoustiques peinent notamment à discriminer les expressions de joie exaltées (elles sont confondues avec les autres émotions fortement activées et avec la colère froide); les caractéristiques vocales perçues, quant à elles, ont du mal à discriminer les expressions d'anxiété qui sont surtout confondues avec les autres expressions faiblement activées.

Paradoxalement, des confusions de cette sorte n'ont pas été rapportées sur le versant du décodage (section 4), alors qu'habituellement les confusions sont envisagées exclusivement à ce niveau (entre les émotions exprimées et les émotions perçues). Les jugements émotionnels ayant été obtenus sur plusieurs échelles d'intensité et la notion de confusion étant directement liée à la procédure de choix forcé, les confusions entre émotions exprimées et émotion perçues n'ont pu être évaluée de manière classique dans cette étude. Toutefois, une forme moins classique de confusion a été rapportée pour les jugements relatifs à l'intensité de peur perçue qui sont statistiquement aussi élevés pour les expressions de désespoir que pour les expressions d'anxiété.

Comparativement aux études habituellement effectuées dans le domaine de la communication vocale des émotions, la présentation des résultats et les méthodes d'analyse restent encore

---

[20] En réalité, l'effet principal des locuteurs sur les paramètres acoustiques a été contrôlé en standardisant, pour chaque locuteur, les valeurs obtenues pour les différents paramètres acoustiques. Une telle standardisation n'a pas été effectuée pour les caractéristiques vocales perçues (pour plus de détails v. section 2 et 3).

relativement classiques dans les sections 2 à 4. En revanche, la section 5 introduit une démarche plutôt inhabituelle. Cette section reprend les données déjà présentées séparément dans les sections 2 à 4 pour les présenter sous un nouvel angle, dans le cadre d'une modélisation complète du processus de communication. La section 5 introduit un certain nombre d'aspects entièrement nouveaux – particulièrement une première présentation des relations entre les caractéristiques vocales (acoustiques et perçues) et les émotions perçues – mais propose également une analyse alternative des relations entre les émotions exprimées et les émotions perçues, ainsi que des relations entre les caractéristiques vocales (acoustiques et perçues) et les émotions exprimées. De plus, deux modèles statistiques différents – bien que fondés tous deux sur une approche corrélationnelle – sont proposés pour représenter les données dans cette section.

Au premier abord, cette multiplication des analyses présentées pour les mêmes données peut sembler redondante, voir génératrice de complexité superflue ou de confusion. A notre avis, les analyses présentées dans la section 5 sont cependant motivées, en premier lieu, par la nécessité de réunir dans un même modèle les deux versants de la communication vocale des émotions, afin de comparer directement la participation des caractéristiques vocales mesurées à l'encodage et au décodage des émotions. D'autre part, une justification supplémentaire peut être invoquée sur le plan des aspects spécifiques des résultats qui sont mis en évidence dans cette section, alors qu'ils sont beaucoup moins apparents dans les sections 2 à 4. Les analyses présentées dans la section 5 ont notamment permis de montrer que les paramètres acoustiques sélectionnés et les caractéristiques vocales jugées ne parviennent pas à rendre compte de la même manière de la communication vocale pour les quatre catégories émotionnelles considérées. Les paramètres acoustiques en particulier ne permettent pas de rendre compte de la relation entre la joie exprimée et la joie perçue, alors qu'ils permettent de médiatiser une grande part de la relation entre la colère exprimée et la colère perçue.

Pour conclure cette discussion relative à la présentation des résultats et aux méthodes statistiques utilisées, nous souhaitons rappeler une notion fondamentale de l'approche corrélationnelle qui est parfois occultée au moment d'interpréter les résultats. Les caractéristiques vocales (acoustiques et perçues) qui corrèlent avec les émotions exprimées et avec les émotions perçues ne sont pas nécessairement exploitées dans le processus de communication. Il est en effet possible que des caractéristiques mesurées n'intervenant pas directement dans le processus de communication soient elles-mêmes corrélées à des caractéristiques non-mesurées qui interviendraient, quant à elles, directement dans le processus de communication. En d'autres termes, il convient de ne pas interpréter les corrélations comme indiquant nécessairement une relation directe entre les aspects mesurés par les variables corrélées.

Cette dernière observation motive notamment la décision prise de ne pas analyser les caractéristiques des contours de F0 (examinées dans la section 6) dans un modèle en lentille représentant le processus de communication. Nous avons en effet observé dans la section 6 que les relations entre la plupart des caractéristiques codées pour les contours de F0 et les émotions exprimées pouvaient être expliquées en faisant appel à des mesures plus élémentaires de la F0 (telles que la moyenne et l'étendue de F0). En conséquence, il nous a semblé plus adéquat de tester la possibilité d'influencer les attributions émotionnelles en manipulant différents aspects des contours de F0 (v. section 6), plutôt qu'en corrélant les caractéristiques des contours de F0 avec les attributions émotionnelles.

7.3 Perspectives

Quelques propositions relatives à des aspects de la recherche présentée qui nécessiteraient d'être améliorés ont été formulées ci-dessus. Nous avons mentionné en particulier la nécessité de développer des modèles qui permettraient de formuler (et de tester) des prédictions relativement aux caractéristiques vocales impliquées dans le processus de communication vocale des émotions. Nous avons aussi évoqué la nécessité de développer des mesures plus adaptées pour évaluer ces caractéristiques vocales (acoustiques et perçues). Ces deux propositions sont partiellement liées. La formulation d'hypothèses spécifiques relativement aux caractéristiques vocales impliquées dans la communication émotionnelle devrait idéalement aboutir au développement de mesures reflétant ces caractéristiques. La question de l'élaboration d'une taxonomie émotionnelle spécifiquement adaptée à l'étude des expressions vocales a également été évoquée ci-dessus. Il semble en particulier nécessaire dans ce domaine d'inclure des états émotionnels plus différenciés que les émotions "fondamentales" habituellement considérées et de développer un système qui permette de caractériser ces différents états, éventuellement en référence à des dimensions sous-jacentes qui permettraient de différencier et de décrire avec précision un grand nombre d'états émotionnels différents.

Dans les paragraphes qui suivent, nous souhaitons élargir le champ de la perspective proposée à quelques problématiques qui n'ont pas été traitées dans la recherche présentée, mais qui se dessinent cependant en filigrane des questions examinées. La question des différences interindividuelles et, plus généralement, la question de la régulation des expressions représentent notamment deux perspectives potentielles de développement pour l'étude de la communication vocale des émotions. Par ailleurs, les expressions vocales mériteraient d'être plus souvent examinées dans différents contextes verbaux et dans différents contextes extra-verbaux. Dans ce domaine, la possibilité et les

avantages de l'étude simultanée des expressions vocales et des expressions faciales sera brièvement évoquée.

7.3.1 Différences interindividuelles

Les différences interindividuelles s'observent aussi bien sur le plan de la production des expressions émotionnelles que sur le plan de leur reconnaissance. Elles ont été considérées par des auteurs provenant de différentes traditions de recherche. Dans le domaine de l'expressivité émotionnelle (encodage), des travaux ont été notamment consacrés aux différences interculturelles (e.g. Matsumoto, 1990), aux différences de genre (e.g. LaFrance & Hecht, 1999, 2000) ou encore aux différences développementales (e.g. Camras, Malatesta, & Izard, 1991). D'autres auteurs qui se sont intéressés aux styles expressifs ont notamment proposé de différencier les individus relativement à leur degré d'expressivité (cf. la distinction entre *internalizers* et *externalizers* proposée par Jones, (1960), en fonction du degré de contrôle qu'ils exercent sur leurs expressions en situation sociale (v. le concept de *self-monitoring* introduit par Snyder (1974) ou encore en fonction de leur "habileté expressive générale" (*successful/unsuccessful communicators*, Friedman & Riggio, 1999). Dans la grande majorité de ces travaux, l'intérêt des auteurs s'est porté essentiellement sur les expressions faciales. Les études consacrées aux différences interindividuelles dans le domaine de l'expression vocale des émotions sont comparativement très peu nombreuses.

Dans le domaine de la perception (reconnaissance) des expressions vocales, des études ont été également consacrées aux différences interculturelles (e.g. Scherer et al., 2001 ; v. aussi Elfenbein & Ambadi, 2002), aux différences de genre (e.g. Hall, Carter, & Horgan, 2000) et aux différences développementales (e.g. De Sonneville et al., 2002). Dans le cadre de l'étude de la perception interpersonnelle (*interpersonal sensitivity*), quelques auteurs se sont intéressés à la relation entre la capacité de reconnaître les expressions émotionnelles et d'autres différences interindividuelles telles que la personnalité ou l'intelligence (v. Rosenthal, Hall, DiMatteo, Rogers, & Archer, 1979; Hall & Bernieri, 2001). Dans ce domaine également, très peu de travaux ont été consacrés spécifiquement aux différences interindividuelles associées à la reconnaissance des émotions communiquées par la voix.

Dans cette thèse, nous avons tenté de contrôler les différences d'expressivité entre les locuteurs/acteurs en utilisant différentes procédures (standardisation des paramètres acoustiques, Anovas à mesures répétées, présentations isolées des expressions produites par différents acteurs dans les études de jugements). A plusieurs reprises toutefois, des différences entre les expressions produites par différents acteurs ont été évoquées. Dans la section 3, les graphiques 8 et 9 mettent par exemple en évidence une variabilité relativement importante des caractéristiques vocales

perçues associées à différentes émotions exprimées pour différents acteurs. Dans la section 4 (graphiques 10 et 12), nous avons constaté des différences relativement à la compétence des différents acteurs à produire des expressions émotionnelles reconnaissables. Les différences entre les auditeurs/juges sur le plan de la reconnaissance des expressions émotionnelles ont été, quant à elles, totalement supprimées par le recours aux jugements moyens. Nous pensons néanmoins qu'il est fort probable que des différences interindividuelles relativement importantes interviennent sur le plan de la capacité à reconnaître les émotions exprimées et, également, sur le plan des "stratégies" utilisées par différents auditeurs pour attribuer une émotion à un locuteur à partir d'une expression vocale. Les différences interindividuelles sur le plan de l'expressivité et sur le plan de la reconnaissance des émotions exprimées représentent, à notre avis, une problématique centrale dans ce domaine d'étude et devraient faire l'objet de plus amples investigations dans des études futures.

7.3.2 Régulation des expressions émotionnelles

La régulation des expressions est un domaine dans lequel les différences interindividuelles sont très marquées (v. ci-dessus). Différentes forme de régulations des expressions sont probablement mises en place par différents individus dans différents contextes (v. Gross, 1998; Gross & John, 1998). Toutefois, au-delà des différences interindividuelles, certaines forme de régulations sont plus ou moins systématiquement imposées par certaines situations sociales. Cet aspect a été évoqué dans l'introduction théorique mais n'a pas été examiné dans le cadre de la recherche présentée. Les régulations appliquées aux émotions peuvent consister notamment à exagérer ou simuler certaines expressions émotionnelles ou au contraire à supprimer ou masquer d'autres expressions.

La notion de régulation est particulièrement importante, si l'on considère qu'elle introduit une complexité supplémentaire sur le plan de la communication non-verbale des émotions. Dans une perspective fonctionnelle, elle implique notamment qu'une variété d'expressions différentes pourraient être éventuellement interprétées par des auditeurs comme correspondant à un même état émotionnel (associé à différentes formes de régulations). En restant dans cette perspective fonctionnelle, les auditeurs seraient capables de reconnaître certaines forme de régulation. Des travaux effectués dans le domaine de la dissimulation des réactions émotionnelles (v. Ekman, O'Sullivan, Friesen, & Scherer, 1991) permettent de penser que certains indicateurs de dissimulation existent et peuvent être utilisés par des observateurs particulièrement compétents dans ce domaine. La plupart des études qui se sont intéressées à la régulation ont examiné la régulation des expressions faciales. Très peu d'études relatives à la régulation des expressions vocales ou à la reconnaissance d'expressions vocales dissimulées ont été effectuées. Ce domaine pourrait toutefois

se révéler riche d'enseignement relativement à la variabilité des expressions susceptibles de communiquer une même information émotionnelle.

7.3.3 Interactions avec des variables contextuelles

Pour clore cette section consacrée aux développements inspirés par la recherche présentée, quelques observations concernant l'influence du contenu verbal des expressions sur la communication des émotions, ainsi que quelques considérations relatives aux interactions entre les expressions vocales et les expressions faciales seront développées ci-dessous.

Influences du contenu verbal des expressions

Dans la recherche présentée ci-dessus, des expressions vocales dépourvues de contenu sémantique (deux séquences de syllabes sans signification) ont été utilisées. Toutefois, l'importance du contenu verbal des expressions, dans des situations plus représentatives d'expériences quotidiennes de communication, a été évoquée à plusieurs reprises. En particulier, nous avons mentionné des résultats obtenus par des études qui se sont intéressée à la participation de l'intonation dans la communication émotionnelle et qui ont mis en évidence des interactions entre le contenu verbal (principalement des aspects syntaxiques) et certains aspects du contour de hauteur relativement aux attributions émotionnelles réalisées par des groupes d'auditeurs (v. sections 1 et 6). Ces travaux indiquent que des impressions émotionnelles peuvent parfois résulter de la transgression d'attentes relativement à des contours intonatifs habituellement associés à différentes structures syntaxiques. Cette proposition peut être élargie également au contenu sémantique des expressions. Bien qu'il n'existe pas d'études dans ce domaine, il est probable que dans des expériences quotidiennes de communication des individus détectent parfois des émotions (ou des attitudes) en se basant sur une *association* entre un contenu sémantique ("ce qui est dit") et des indicateurs vocaux non-verbaux ("la manière dont c'est dit").

Un point de développement potentiel pour l'étude de la communication vocale des émotions serait en conséquence de formuler et de tester un plus grand nombre d'hypothèses relativement aux interactions potentielles entre différents aspects verbaux (syntaxiques, sémantiques) et non-verbaux.

Relations entre expressions faciales et vocales

Les expressions vocales et les expressions faciales sont le plus souvent étudiées par différents auteurs dans des études indépendantes. Il existe toutefois quelques exceptions, en particulier dans le domaine de l'étude de la reconnaissance des expressions, où certains auteurs ont tenté d'évaluer l'influence respective de différentes modalités de communication. Dans ce domaine, il a été parfois

suggéré que les expressions faciales jouent un rôle plus important que les expressions vocales dans la communication émotionnelle. Une étude d'Ekman, Friesen, O'Sullivan, & Scherer (1980) contredit cette proposition. Dans cette étude, des jugements relatifs à différentes caractéristiques liées à la personnalité et aux émotions de plusieurs personnes placées dans deux contextes différents (un contexte de communication "spontanée" et un contexte de dissimulation) ont été obtenus sur la base d'enregistrements audio-vidéo. Des jugements pour les mêmes caractéristiques obtenus uniquement sur la base des enregistrements audio (incluant des aspects verbaux et non-verbaux), des jugements obtenus uniquement sur la base de l'observation des expressions faciales (sans l'information relative aux expressions verbales/vocales) et des jugements obtenus uniquement sur la base de l'observation des expressions corporelles (sans l'information relative au visage et aux expressions verbales/vocales) ont été corrélés aux jugements obtenus pour les enregistrements comprenant toutes les modalités.

Les résultats de cette étude indiquent que l'avantage des jugements basés sur l'une ou l'autre modalité dépend du type de jugement effectué et de la situation de communication considérée. Les jugements basés sur les expressions vocales/verbales ont par exemple permis de réaliser une meilleure prédiction des jugements globaux dans le contexte de dissimulation. Sur un plan plus général, les résultats de ces études indiquent que les stratégies de perception/reconnaissance peuvent varier en fonction du contexte et que des indicateurs différents (dans ce cas des indicateurs faciaux, posturaux ou vocaux/verbaux) auront une influence différente dans différents contextes. De plus, les corrélations multiples entre les jugements obtenus pour les modalités isolées et les jugements obtenus lorsque toutes les modalités sont présentées simultanément sont parfois relativement faibles. Les auteurs de cette étude ont en conclu qu'une partie des jugements réalisés lorsque toutes les modalités sont présentées simultanément découlent d'interactions entre les informations disponibles dans les différentes modalités.

D'autre part, certaines revue de la littérature ont récemment suggéré que de manière générale certaines émotions (en particulier la joie et le dégoût) seraient préférentiellement communiquées par les expressions faciales, alors que d'autres expressions seraient préférentiellement communiquées par les expressions vocales (v. Scherer, 2003).

Dans une autre perspective encore, un autre type d'interaction entre les expressions faciales et les expressions vocales a été mis en avant. Quelques auteurs se sont récemment intéressés aux indicateurs vocaux qui reflètent des modifications de certaines configurations musculaires qui sont par ailleurs apparentes sur le plan facial. En particulier, la possibilité de détecter la présence de sourires sur le plan des expressions vocales a retenu l'attention de ces auteurs. Dans le même sens,

certains aspects des expressions vocales/verbales – par exemple la fluidité et la qualité de l'articulation ou encore le rythme de parole – peuvent très probablement être "lues" sur les expressions faciales (v. de Gelder, Vrommen, & Pourtois, 1999). En d'autres termes, certains aspects des expressions faciales peuvent être entendus et certains aspects des expressions vocales peuvent être vus. Ces deux modalités de communications ne sont donc jamais totalement indépendantes.

Ces quelques exemples d'interactions possibles entre les expressions vocales et faciales indiquent que les motivations pour le développement de l'étude simultanée des expressions vocales et faciales ne manquent pas. Dans le même sens, nous pensons que les approches et les résultats présentés dans cette thèse confirment l'intérêt d'étudier le processus de communication émotionnelle en associant les études de production (encodage) et les études de perception (décodage), dans le domaine spécifique de la recherche sur les expressions vocales et, également, dans le domaine plus général de la recherche sur les expressions émotionnelles (vocales et faciales).

Ainsi que nous l'avons mentionné dans les premières pages de cette thèse, l'étude des expressions émotionnelles – en particulier l'étude des expressions vocales – a connu récemment (et connaît encore) un très fort développement. L'accroissement du nombre de travaux consacrés à ce domaine de recherche a été (et est encore) fortement motivé par le développement des technologies de communication qui représentent actuellement le champ d'application privilégié de ce domaine d'étude. Des applications telles que la reconnaissance automatique de la parole, la synthèse de la parole, ou encore la reconnaissance automatique des locuteurs bénéficieraient incontestablement des progrès qui pourraient être réalisés dans ce domaine d'étude. Dans un futur proche, nous formulons l'espoir que les paradigmes que nous avons développés et les résultats que nous avons présentés dans cette thèse et, plus généralement, les approches développées par des psychologues spécialisés dans le domaine de l'étude des expressions faciales et vocales puissent être intégrés aux développements de ces nouvelles technologies de communication.

8 Bibliographie

Albright, L., & Malloy, T. E. (2001). Brunswik's theoretical and methodological contributions to research in interpersonal perception. In K. R. Hammond & T. R. Stewart (Eds.), *The essential Brunswik: Beginnings, explications, applications* (pp. 328-332). New York: Oxford University Press.

Bachorowski, J. A., & Owren, M. J. (1995). Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science, 6*(4), 219-224.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614-636.

Bergmann, G., Goldbeck, T., & Scherer, K. R. (1988). Emotionale Eindruckswirkung von prosodischen Sprechmerkmalen. *Zeitschrift fuer Experimentelle und Angewandte Psychologie, 35*(2), 167-200.

Bernieri, F. J., & Gillis, J. S. (2001). Judging rapport: employing Brunswik's lens model to study interpersonal sensitivity. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 67-88). Mahwah, NJ: Lawrence Erlbaum Associates.

Boersma, P., & Weenink, D. J. M. (1996). *Praat, a system for doing phonetics by computer, version 3.4* (132). Amsterdam: Institute of Phonetic Sciences of the University of Amsterdam.

Breitenstein, C., Van Lancker, D., & Daum, I. (2001). The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition and Emotion, 15*(1), 57-79.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.

Burkhardt, F. (2001). *Simulation emotionaler Sprechweise mit Sprachsystemen*. Aachen: Shaker Verlag.

Camras, L. A., Malatesta, C., & Izard, C. E. (1991). The development of facial expressions in infancy. In R. S. Feldman & B. Rimé (Eds.), *Fundamentals of nonverbal behavior* (pp. 73-105). Cambridge: Cambridge University Press.

Chung, S.-J. (2000). *L'expression et la perception de l'émotion extraite de la parole spontannée: évidences du coréen et de l'anglais.* Université de la Sorbonne Nouvelle, Paris.

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

Cowie, R. (2000). *Describing the emotional states expressed in speech.* Paper presented at the Proceedings of the ISCA workshop on speech and emotion, Newcastle, Northern Ireland.

Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.

Crystal, D. (1969). *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

Darwin, C., & Ekman, P. (1998). *The expression of the emotions in man and animals* (3rd ed.). London: Oxford University Press.

Davitz, J. R. (1964). *The communication of emotional meaning*. Oxford, England: McGraw Hill.

de Gelder, B., Vrommen, J., & Pourtois, G. (1999). Seeing cries and hearing smiles: Crossmodal perception of emotional expressions. In G. Aschersleben & T. Bachmann (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 425-438). Amsterdam, Netherlands: Elsevier Science Publishers.

De Sonneville, L. M. J., Vershoor, C. A., Njiokiktjien, C., Veld, V. O. h., Toorenaar, N., & Vranken, M. (2002). Facial identity and facial emotions: Speed, accuracy and processing strategies in children and adults. *Journal of Clinical and Experimental Neuropsychology, 24*(2), 200-213.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*, 169-200.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1*, 49-98.

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: A technique for the measurement of facial movement*. Palo Alto: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., O'Sullivan, M., & Scherer, K. (1980). Relative importance of face, body, and speech in judgments of personality and affect. *Journal of Personality and Social Psychology, 38*(2), 270-277.

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science, 221*, 1208-1210.

Ekman, P., O'Sullivan, M., Friesen, W., & Scherer, K. R. (1991). Face, voice, and body in detecting deceit. *Journal of Nonverbal Behavior, 15*, 125-135.

Eldred, S. H., & Price, D. B. (1958). A linguistic evaluation of feeling states in psychotherapy. *Psychiatry: Journal for the Study of Interpersonal Processes, 21*, 115-121.

Elfenbein, H. A., & Ambadi, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128*, 203-235.

Fairbanks, G., & Pronovost, W. (1939). An experimental study of the pitch characteristics of the voice during the expression of emotions. *Speech Monographs, 6*, 87-104.

Fernald, A. (1991). Prosody in speech to children: Prelinguistic and linguistic functions. In R. Vasta (Ed.), *Annals of child development* (pp. 43-80). London: Jessica Kingsley Publishers.

Fernald, A. (1992). Meaningful melodies in mothers' speech to infants. In H. Papousek & U. Juergens (Eds.), *Nonverbal vocal communication: Comparative and developmental approaches* (pp. 262-282). Cambridge: Cambridge University Press.

Fernald, A. (1993). Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages. *Child Development, 64*(3), 657-674.

Fonagy, I., & Magdics, K. (1963). Emotional patterns in intonation and music. *Zeitschrift für Phonetik, 16*, 293-326.

Frick, R. W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin, 97*(3), 412-429.

Frick, R. W. (1986). The prosodic expression of anger: Differentiating threat and frustration. *Aggressive Behavior, 12*(2), 121-128.

Friedman, H. S., & Riggio, R. E. (1999). Individual differences in ability to encode complex affects. *Personality and Individual Differences, 27*(1), 181-194.

Frijda, N. H., Ortony, A., Sonnemans, J., & Clore, G. L. (1992). The complexity of intensity: Issues concerning the structure of emotion intensity. In M. S. Clark (Ed.), *Emotion. Review of personality and social psychology* (Vol. 13, pp. 60-89). Thousand Oaks, CA, US: Sage Publications.

Fujisaki, H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In O. Fujimura (Ed.), *Vocal physiology: voice production, mechanisms and functions* (pp. 347-355). New York: Raven.

Fujisaki, H. (1997). Prosody, models and spontaneous speech. In Y. Sagisaka, N. Campbell & N. Higuchi (Eds.), *Computing prosody. Computational models for processing spontaneous speech* (pp. 27-42). New York: Springer.

Fulchner, J. A. (1991). Vocal affect expression as an indicator of affective response. *Behaviour Research Methods, Instruments, & Computers, 23*, 306-313.

Gifford, R. (1994). A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behavior. *Journal of Personality and Social Psychology, 66*(2), 398-412.

Granqvist, S. (1996). *Enhancements to the visual analogue scale* (4). Stockholm: Speech, Music and Hearing - Quaterly Progress and Status Report, KTH.

Gray, J. A. (1987). *The psychology of fear and stress* (2nd ed.). Cambridge: Cambridge University Press.

Green, R. S., & Cliff, N. (1975). Multidimensional comparisons of structures of vocally and facially expressed emotion. *Perception and Psychophysics, 17*(5), 429-438.

Gross, J. J. (1998). Antecedent- and response-focused emotion regulation: divergent consequences for experience, expression, and physiology. *Journal of personality and social psychology, 74*, 224-237.

Gross, J. J., & John, O. P. (1998). Mapping the domain of expressivity: Multimethod evidence for a hierarchical model. *Journal of Personality and Social Psychology, 74*, 170-191.

Guerrero, L. K., Andersen, P. A., & Trost, M. R. (1998). Communication and emotion: Basic concepts and approaches. In P. A. Andersen & L. K. Guerrero (Eds.), *Handbook of communication and emotion: Research, theory, applications and contexts* (pp. 3-27). San Diego: Academic Press.

Hall, J. A., & Bernieri, F. J. (2001). *Interpersonal sensitivity: Theory and measurement*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hall, J. A., Carter, J. D., & Horgan, T. G. (2000). Gender differences in nonverbal communication of emotion. In A. H. Fisher (Ed.), *Gender and emotion: Social psychological perspectives*. Cambridge: Cambridge University Press.

Halliday, M. A. K. (1970). *A course in spoken English: Intonation*. Oxford: Oxford University Press.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review, 62*, 255-262.

Hammond, K. R. (2001). Expansion of Egon Brunswik's psychology, 1955-1995. In K. R. Hammond & T. R. Stewart (Eds.), *The essential Brunswik: Beginnings, explications, applications* (pp. 464-478). New York: Oxford University Press.

Hammond, K. R., & Stewart, T. R. (2001). *The essential Brunswik: Beginnings, explications, applications*. New York: Oxford University Press.

Hammond, K. R., Wilkins, M. M., & Todd, F. J. (1966). A research paradigm for the study of interpersonal learning. *Psychological Bulletin, 65*(4), 221-232.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1994). *Emotional contagion*. Cambridge: Cambridge University Press.

Heilman, K. M., Bowers, D., Speedie, L., & Coslett, H. B. (1984). Comprehension of affective and nonaffective prosody. *Neurology, 34*(7), 917-921.

Hess, U., & Blairy, S. (2001). Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. *International Journal of Psychophysiology, 40*(2), 129-141.

Holzworth, R. J. (2001). Judgment analysis. In K. R. Hammond & T. R. Stewart (Eds.), *The essential Brunswik: Beginnings, explications, applications* (pp. 324-327). New York: Oxford University Press.

Hursh, C. J., Hammond, K. R., & Hursh, J. L. (1964). Some methodological considerations in multiple-cue probability learning studies. *Psychological Review, 71*, 42-60.

James, W. (1884). What is an emotion? *Mind, 9*, 188-205.

Johnstone, T. (2001). *The effect of emotion on voice production and speech acoustics.* University of Western Australia, Perth.

Johnstone, T., & Scherer, K. R. (2000). Vocal communciation of emotion. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of Emotions* (2nd ed., pp. 220-235). New York: Guilford Press.

Johnstone, T., Van Reekum, C. M., & Scherer, K. R. (2001). Vocal expression correlates of appraisal processes. In K. R. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research.* (pp. 271-284). New York: Oxford University Press.

Jones, H. G. (1960). The longitudinal method in the study of personality. In I. Iscoe & H. W. Stevenson (Eds.), *Personality development in children*. Chicago: University of Chicago Press.

Juslin, P. N. (1998). *A functionalistic perspective on emotional communication in music performance (doctoral dissertation)*. Uppsala, Sweden: Uppsala University Library.

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion, 1*(4), 381-412.

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

Katz, G. S. (1998). *Emotional speech: A quantitative study of vocal acoustics in emotional expression.* University of Pittsburgh, Pittsburgh, PA.

Kreiman, J. (1998). Listening to voices: Theory and practice in voice perception research. In K. Johnson & J. W. Mulenmix (Eds.), *Talker variability in Speech Processing*. New York: Academic Press.

Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *Journal of the Acoustical Society of America, 104*(3), 1598-1608.

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35*(3), 512-520.

Kutik, E. J., Cooper, W. E., & Boyce, S. (1983). Declination of fundamental frequency in speaker's production of parenthetical and main clauses. *Journal of the Acoustical Society of America, 73*, 1731 - 1738.

Lacheret-Dujour, A., & Beaugendre, F. (1999). *La prosodie de français*. Paris: Editions du CNRS.

Ladd, D. R. (1983a). Peak features and overall slope. In A. Cutler & D. R. Ladd (Eds.), *Prosody: models and measurements* (pp. 39-52). New York: Springer-Verlag.

Ladd, D. R. (1983b). Phonological features of intonational peaks. *Language, 59*, 721-759.

Ladd, D. R., Silverman, K., Tolkmitt, F., Bergmann, G., & Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signalling speaker affect. *Journal of the Acoustical Society of America, 78*, 435-444.

LaFrance, M., & Hecht, M. A. (1999). Option or obligation to smile: The effects of power and gender on facial expression. In P. Philippot & R. S. Feldman (Eds.), *The social context of nonverbal behavior* (pp. 45-70). Cambridge: Cambridge University Press.

LaFrance, M., & Hecht, M. A. (2000). Gender and smiling: A meta-analysis. In A. H. Fischer (Ed.), *Gender and emotion: Social psychological perspectives* (pp. 118-142). New York: Cambridge University Press.

Laukka, P., Juslin, P. N., & Bresin, R. (submitted). A dimensional approach to vocal expression of emotion. *Cognition and Emotion*.

Laukkanen, A.-M., Vilkman, E., Alku, P., & Oksanen, H. (1996). Physical variations related to stress and emotional state: A preliminary study. *Journal of Phonetics, 24*, 313-335.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

Léon, P. R., & Martin, P. (1970). *Prolégomènes à l'étude des structures intonatives*. Montréal: Marcel Didier.

Lieberman, P., & Michaels, S. B. (1962). Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *Journal of the Acoustical Society of America, 34*(7), 922-927.

Lipps, T. (1903). Die Einfühlung. In T. Lipps (Ed.), *Leitfaden der Psychologie* (pp. 187-201). Leipzig: Verlag von Wilhelm Engelmann.

Matsumoto, D. (1990). Cultural similarities and differences in display rules. *Motivation and Emotion, 14*(3), 195-214.

Mertens, P. (1987). *L'intonation du français. De la description linguistique à la reconnaissance automatique*. KU Leuven, Belgium.

Morel, M., & Bänziger, T. (2004). Le rôle de l'intonation dans la communication vocale des émotions : test par la synthèse. *Cahiers de l'Institut de Linguistique de Louvain (CILL), 30*, 207-232.

Morel, M., & Lacheret-Dujour, A. (2001). Le logiciel de synthèse vocale Kali : de la conception à la mise en œuvre. In C. d'Alessandro (Ed.), *Traitement Automatique des Langues* (Vol. 42, pp. 193-221). Paris: Hermès.

Mozziconacci, S. J. (1998). *Speech variability and emotion: production and perception.* Technische Universiteit Eindhoven, Eindhoven.

Murray, I. R., Arnott, J. L., & Rohwer, E. A. (1996). Emotional stress in synthetic speech: Progress and future directions. *Speech Communication, 20*(1-2), 85-91.

O'Connor, J. D., & Arnold, G. F. (1973). *Intonation of colloquial English* (2nd ed.). London: Longman.

Osgood, C. E. (1966). Dimensionality of the semantic space for communication via facial expressions. *Scandinavian Journal of Psychology, 7*, 1-30.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Pakosz, M. (1983). Attitudinal judgments in intonation: Some evidence for a theory. *Journal of Psycholinguistic Research, 12*(3), 311-326.

Papousek, M., Papousek, H., & Symmes, D. (1991). The meanings of melodies in motherese in tone and stress languages. *Infant Behavior and Development, 14*(4), 415-440.

Patterson, D., & Ladd, D. R. (1999). *Pitch range modelling: linguistic dimensions of variation.* Paper presented at the XIVth International Congress of Phonetic Sciences (ICPhS), San Francisco.

Pell, M. D. (1998). Recognition of prosody following unilateral brain lesion: Influence of functional and structural attributes of prosodic contours. *Neuropsychologia, 36*(8), 701-715.

Pell, M. D. (2001). Influence of emotion and focus location on prosody in matched statements and questions. *Journal of the Acoustical Society of America, 109*(4), 1668-1680.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation.* Massachusetts Institute of Technology.

Pittam, J., & Scherer, K. R. (1993). Vocal expression and communication of emotion. In M. Lewis & J. M. Haviland (Eds.), *Handbook of emotions* (pp. 185-197). New York: Guilford Press.

Plutchik, R. (1962). *The emotions: Facts, theories and a new model*. New York: Random House.

Reissland, N., Shepherd, J., & Cowie, L. (2002). The melody of surprise: Maternal surprise, vocalizations during play with her infant. *Infant and Child Development, 11*(3), 271-278.

Reynolds, D. A. J., & Gifford, R. (2001). The sounds and the sights of intelligence: A Lens Model channel analysis. *Personality and Social Psychology Bulletin, 27*(2), 187-200.

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication. The PONS test.* Baltimore: The Johns Hopkins University Press.

Rosenthal, R., & Rubin, D. B. (1989). Effect size estimation for one-sample multiple-choice-type data: Design, analysis and meta-analysis. *Psychological Bulletin, 106*, 332-337.

Ross, E. D. (1981). The aprosodias: Functional-anatomic organization of the affective components of language in the right hemisphere. *Annals of Neurology, 38*(561-589).

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161-1178.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin, 115*, 102-141.

Russell, J. A., & Feldman Barrett, L. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology, 76*(5), 805-819.

Safer, M. A., & Leventhal, H. (1977). Ear differences in evaluating emotional tones of voice and verbal content. *Journal of Experimental Psychology: Human Perception and Performance, 3*(1), 75-82.

Sangsue, J., Siegwart, H., Grosjean, M., Cosnier, J., Cornut, J., & Scherer, K. R. (1997). *Développement d'un questionaire d'évaluation subjective de la qualité de la voix et de la parole, QEV*. Geneva Studies in Emotion and Communication, 11 (1). Retrieved, 2003, from the World Wide Web: <http://www.unige.ch/fapse/emotion/genstudies/genstudies.html>

Schachter, S., & Singer, J. E. (1962). Cognitive, social and physiological determinants of emotional state. *Psychological Review, 69*, 379-399.

Scherer, K., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross Cultural Psychology, 32*(1), 76-92.

Scherer, K. R. (1978). Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology, 8*, 467-487.

Scherer, K. R. (1982). Methods of research on vocal communication: Paradigms and parameters. In K. R. Scherer & P. Ekman (Eds.), *Handbook of methods in nonverbal behavior research* (pp. 136-198). Cambridge: Cambridge University Press.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99*(2), 143-165.

Scherer, K. R. (1989). Vocal correlates of emotion. In H. Wagner & A. Manstead (Eds.), *Handbook of psychophysiology: Emotion and social behavior* (pp. 165-197). London: Wiley.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*, 227-256.

Scherer, K. R., Feldstein, S., Bond, R. N., & Rosenthal, R. (1985). Vocal cues to deception: A comparative channel approach. *Journal of Psycholinguistic Research, 14*(4), 409-425.

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In R. J. Davidson & H. Goldsmith & K. R. Scherer (Eds.), *Handbook of the affective sciences* (pp. 433-456). Oxford: Oxford University Press.

Scherer, K. R., Ladd, D. R., & Silverman, K. E. A. (1984). Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America, 76*, 1346-1356.

Scherer, K. R., & Oshinsky, J. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion, 1*, 331-346.

Schlossberg, H. (1954). Three dimensions of emotion. *Psychological Review, 61*, 81-88.

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology, 52*, 1061-1086.

Silverman, K., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C., Price, P., & Hirschberg, J. (1992). *ToBI: A standard scheme for labeling prosody.* Paper presented at the International Conference on Spoken Language Processing (ICSLP), Beijing, China.

Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology, 30*, 526-537.

Sobin, C., & Alpert, M. (1999). Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy. *Journal of Psycholinguistic Research, 28*, 347-365.

Stewart, T. R. (2001). The Lens Model Equation. In K. R. Hammond & T. R. Stewart (Eds.), *The essential Brunswik: Beginnings, explications, applications* (pp. 357-362). New York: Oxford University Press.

Stewart, T. R., & Lusk, C. M. (1994). Seven components of judgmental forecasting skill: Implications for research and the improvement of forecasts. *Journal of Forecasting, 13*, 575-599.

Stewart, T. R., Moninger, W. R., Heideman, K. F., & Reagan-Cirincione, P. (1992). Effects of improved information on the components of skill in weather forecasting. *Organizational Behaviour and Human Decision Processes, 53*, 107-134.

Sundberg, J. (1987). *The science of the singing voice*. DeKalb: Northern Illinois University Press.

Sundberg, J., Iwarsson, J., & Hagegård, H. (1995). A singer's expression of emotions in sung performance. In O. Fujimura & M. Hirano (Eds.), *Vocal fold physiology, voice quality control* (pp. 217-231). San Diego: Singular.

Syrdal, A. K., & McGory, J. (2000). *Inter-transcriber reliability of ToBI prosodic labeling.* Paper presented at the International Conference on Spoken Language Processing (ICSLP), Beijing, China.

t'Hart, J., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation*. Cambridge: Cambridge University Press.

Tolkmitt, F., Helfrich, H., Standke, R., & Scherer, K. R. (1982). Vocal indicators of psychiatric treatment effects in depressives and schizophrenics. *Journal of Communication Disorders, 15*(3), 209-222.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursh, Hammond, & Hursh and by Hammond, Hursh, & Todd. *Psychological Review, 71*, 528-530.

Uldall, E. (1964). Dimensions of meaning in intonation. In D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. C. Scott & J. L. M. Trim (Eds.), *In honour of Daniel Jones: papers contributed on the occasion of his eightieth birthday, 12 september 1961* (pp. 271-279). London: Longman.

van Bezooijen, R. A. (1984). *Characteristics and recognizability of vocal expressions of emotion*. Dordrecht, The Netherlands: Foris Publications.

van Lancker, D., & Sidtis, J. J. (1992). The identification of affective-prosodic stimuli by left- and right-hemisphere-damaged subjects: All errors are not created equal. *Journal of Speech and Hearing Research, 35*(5), 963-970.

Wallbott, H. G. (1991). Recognition of emotion from facial expression via imitation? Some evidence for an old theory. *British Journal of Social Psychology, 30*, 207-219.

Wightman, C. W. (2002). *ToBI or not ToBI.* Paper presented at the International Conference on Speech Prosody 2002, Aix-en-Provence, France.

Wigton, R. S. (1996). Social judgement theory and medical judgement. *Thinking and reasoning, 2*, 175-190.

Wigton, R. S. (2001). Brunswik and medical science. In K. R. Hammond & T. R. Stewart (Eds.), *The essential Brunswik: Beginnings, explications, applications* (pp. 378-379). New York: Oxford University Press.

Williams, C. E., & Stevens, K. N. (1969). On determining the emotional state of pilots during flights: An exploratory study. *Aerospace Medicine, 40*, 1369-1372.

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America, 52*, 1238-1250.

Wundt, W. (1897). *Outlines of psychology*. New York: Gustav E. Stechert.