Prosodie de l'émotion :
étude de l'encodage et du décodage

T. Bänziger, D. Grandjean, P. J. Bernard,
G. Klasmeyer & K. R. Scherer
FAPSE, Université de Genève
<Tanja.Banziger@pse.unige.ch>

1. Introduction

L'importance de la communication non-verbale des émotions dans le discours a retenu l'attention des rhétoriciens dès l'Antiquité. Les premières études empiriques relatives à la communication émotionnelle vocale remontent au début du 20ème siècle, mais la recherche dans ce domaine a connu un développement considérable surtout durant les quatre dernières décennies et on observe encore actuellement un intérêt croissant pour les fonctions pragmatiques de la parole. Trois domaines de recherche – la psychologie de l'émotion, la linguistique et les technologies de la parole – ont particulièrement favorisé et favorisent encore l'étude de la communication vocale des attitudes, des humeurs et des émotions. En psychologie de l'émotion, un intérêt accru pour l'étude plus générale de l'expression émotionnelle s'est traduit par un accroissement des études sur la communication vocale des émotions. En linguistique pragmatique, un intérêt s'est développé pour les formes intonatives et leurs fonctions, y compris la fonction expressive (émotionnelle). Enfin, dans le cadre du développement des technologies de la parole, il est apparu qu'il est nécessaire de prendre en compte les modifications vocales associées aux émotions et aux attitudes des locuteurs afin d'améliorer d'une part la performance des systèmes de reconnaissance de la parole et du locuteur et d'autre part l'acceptabilité des systèmes de synthèse vocale.

Ce chapitre présente un survol de l'état actuel de la question. Dans le cadre d'une introduction générale relative à l'étude de la communication vocale des émotions, l'importance de l'étude plus spécifique de l'intonation des expressions émotionnelles est discutée sur le plan théorique et relativement à quelques résultats empiriques fournis par la littérature. Des approches actuellement mises en œuvre par notre groupe de recherche à la

faculté de Psychologie de l'Université de Genève afin d'étudier la prosodie émotionnelle sont présentées.[1]

2. Etude de la communication vocale des émotions

Une revue exhaustive des nombreux travaux consacrés à la communication vocale des émotions est hors de la portée de ce chapitre (v. Johnstone & Scherer 2000 ; Scherer, Johnstone, & Klasmeyer 2002 pour des revues récentes de la littérature). Seules les tendances générales qui émergent de ce courant de recherches seront décrites dans ce qui suit.

Les recherches réalisées peuvent être classées en deux catégories en fonction de leur centre d'intérêt principal. Certaines études se centrent sur les processus d'encodage de l'émotion dans la voix. Ces études s'efforcent de décrire l'effet de différents états affectifs sur un ensemble de paramètres acoustiques. Plus rarement, elles peuvent proposer des explications concernant la manière dont les émotions produisent ces effets. D'autres études se centrent sur les processus de décodage. Ces études s'intéressent principalement à mettre en évidence la capacité des individus à reconnaître 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 descriptions acoustiques des expressions aux attributions émotionnelles sont beaucoup plus rares.

3. Etude de l'encodage de l'émotion dans la voix

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 des profils acoustiques, généralement pour un nombre restreint d'émotions. Les expressions émotionnelles étudiées dans ce domaine sont parfois enregistrées dans des contextes naturellement inducteurs d'émotions ou dans des contextes d'induction émotionnelle en laboratoire. L'utilisation d'expressions émotionnelles simulées par des acteurs est souvent préférée aux deux méthodes précédentes. 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. D'autre part, les

[1] Ces approches ont été développées dans le cadre et avec le soutien du projet plurifacultaire « Prosodie » de l'Université de Genève.

expressions enregistrées en situation d'induction émotionnelle correspondent à des modifications légères de l'état affectif des individus et ne sont pas aussi prononcées que les expressions obtenues en situation naturelle ou en faisant appel à des acteurs. L'utilisation très répandue des expressions simulées par les acteurs a été parfois critiquée relativement au fait que ces expressions ne correspondraient pas ou peu à des expressions émotionnelles « authentiques ». Il est en réalité 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 et il est possible qu'une partie de la composante expressive associée normalement à la réaction physiologique émotionnelle, soit absente des enregistrements produits par les acteurs.

La fréquence fondamentale (F0), l'intensité (amplitude) et la durée de différents segments[2] des expressions sont les paramètres acoustiques les plus fréquemment mesurés. Pour ces paramètres, les tendances centrales (moyennes, médianes) et la variabilité globale (écarts-types, écarts entre minima et maxima) sont en général calculées pour chaque expression étudiée, puis pour chaque émotion exprimée. Un certain nombre d'études incluent l'évaluation de paramètres spectraux (distribution relative de l'énergie dans différentes bandes spectrales, pente spectrale, analyse de formants)[3]. L'inclusion des paramètres spectraux dans les études d'encodage s'est récemment accrue. Mais à ce jour, il existe davantage de données concernant les paramètres de fréquence fondamentale, d'intensité et de durée que de données relatives aux paramètres spectraux.

La comparaison des profils acoustiques établis pour des émotions étudiées par différents chercheurs révèle un consensus relativement faible. Les profils spécifiques établis pour chaque émotion étudiée sont rarement reproduits d'une étude à l'autre. Globalement, le consensus semble se limiter au degré d'activation associé à l'état émotionnel exprimé. 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

---

[2] Les expressions sont le plus souvent segmentées en parties voisées, non voisées et silencieuses.

[3] Un inventaire plus complet de paramètres acoustiques susceptibles de différencier les émotions sera présenté dans la suite de ce chapitre.

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 ou l'ennui – présentent une diminution des valeurs de F0 et d'intensité, ainsi qu'une augmentation de la durée de différents segments (v. Johnstone et Scherer 2000 pour une revue plus détaillée). Ce faible consensus concernant les caractéristiques acoustiques correspondant à différentes émotions pourrait être lié à l'utilisation de différentes procédures[4] et de différentes définitions des états émotionnels considérés qui ne seraient dès lors pas comparables. L'absence de profils différenciés pourrait également être due aux mesures acoustiques utilisées. 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.

Afin de définir quels paramètres acoustiques pourraient être systématiquement affectés par l'état émotionnel du locuteur, il semble nécessaire de développer un modèle détaillé concernant les processus d'encodage. Scherer (1986) a développé un tel modèle et formulé des prédictions concernant les effets sur la voix attendus pour différents changements de l'état émotionnel d'un locuteur. Selon ce modèle théorique l'état émotionnel d'un locuteur affecte l'ensemble du système de production vocale. Le fonctionnement des appareils respiratoire, phonatoire et articulatoire est modifié par les changements cognitifs, autonomes et somatiques associés à la réaction émotionnelle. Sur le plan acoustique les paramètres de fréquence fondamentale, d'intensité et de durée sont liés essentiellement à des modifications qui surviennent au niveau de la respiration et de la phonation. Les changements qui affectent la partie supérieure du conduit vocal où se réalise l'articulation ne sont pas directement reflétés dans ces mesures. Plus généralement, sur le plan perceptif, le modèle de Scherer stipule que les changements physiologiques et somatiques associés aux réactions émotionnelles vont affecter le timbre vocal. Les changements de timbre vocal sont attribuables d'une part au mode de phonation (v. Laver 1980) et d'autre part à la configuration de la partie supérieure du conduit vocal. Dans cette perspective, il est important d'inclure dans les études d'encodage davantage de paramètres reflétant les modifications du timbre vocal. Dans une étude publiée en 1996, Banse et

[4] L'utilisation d'expressions simulées par des acteurs relativement à l'utilisation d'expressions enregistrées en contexte réel et en contexte d'induction émotionnelle pourrait notamment expliquer une partie des différences entre les résultats rapportés par différents chercheurs.

Scherer ont intégré un grand nombre de mesures spectrales, ils ont pu de la sorte établir des profils différenciés pour plusieurs états émotionnels différents.

3.1. Perspectives méthodologiques pour les recherches actuelles et futures

Les techniques d'analyse de signaux permettent aujourd'hui de réaliser l'extraction de nombreux paramètres acoustiques. Dans ce contexte, il est plus important que jamais de trouver un accord sur les mesures qu'il convient de réaliser afin de représenter adéquatement l'encodage de l'émotion dans la voix. Les mesures choisies devraient permettre de différencier les états émotionnels exprimés (induits ou simulés par des acteurs) au-delà de leur dimension d'activation. Afin d'évaluer la capacité des mesures choisies à réaliser cette différenciation, il est nécessaire que les mêmes mesures soient appliquée à de grands/nombreux corpus de parole émotionnelle.

A titre d'exemple, nous présentons ci-dessous un ensemble d'analyses automatisées qui ont été développées par l'un des auteurs de cette contribution (G. Klasmeyer) à la Faculté de Psychologie de l'Université de Genève (pour plus de détails v. Klasmeyer 2000). Les différentes mesures décrites ont été choisies sur la base d'hypothèses concernant les caractéristiques vocales qui sont théoriquement affectées par l'état émotionnel des locuteurs et sur la base de recherches empiriques dans ce domaine (Banse & Scherer 1996 ; Klasmeyer 1999). L'automatisation de l'extraction des paramètres étant nécessaire lorsque l'on souhaite effectuer des mesures pour de grands/nombreux corpus d'expressions émotionnelles, les mesures sélectionnées n'incluent pas de mesures phonétiques complexes qui sont nécessairement réalisées manuellement, mais uniquement des mesures qui peuvent être effectuées (semi-)automatiquement.

La première tâche de l'analyse acoustique consiste à segmenter le signal acoustique en différentes unités phonétiques. Cette segmentation est effectuée en plusieurs étapes présentées dans le **tableau 1**.

Etape	Description
1. Détection des silences avant et après un énoncé	Un calcul du contour d'énergie du signal (fréquence d'échantillonnage : 16 kHz) est effectué avec une fenêtre de 300 points et un pas de 30 points. Toutes les valeurs d'énergie en dessous de 1% de l'énergie maximum absolue du signal sont considérées comme « silencieuses ». Les silences situés avant et après l'énoncé sont supprimés.
2. Vérification par l'oreille humaine	Les limites détectées automatiquement sont contrôlées par un auditeur humain qui signale également les éventuels bruits indésirables ou parties manquantes.
3. Détection des signaux et des pauses	Les valeurs d'énergie en dessous de 1% de l'énergie maximum absolue à l'intérieur de l'énoncé sont considérées comme étant des pauses. Toutes les parties contenant plus d'énergie acoustique sont traitées en tant que signal.
4. Segments voisés et non-voisés	Une routine de détection de périodicité est utilisée pour calculer la probabilité de voisement.
5. Détection automatique de « syllabes »	Un détecteur de « syllabes » rudimentaire, basé sur la stylisation du contour d'énergie, permet une détection approximative des syllabes CVC. Les segments du contour d'énergie considérés comme « syllabes » correspondent à une augmentation de l'énergie suivie d'une diminution de l'énergie (avec ou sans segment stable entre la montée et la descente) d'une durée minimum de 100 ms.
6. Détection des « voyelles accentuées »[5]	Les critères de détections pour une « voyelle accentuée » sont: 1. un segment stable de plus de 60 ms dans une « syllabe » et 2. des valeurs d'énergie au-dessus de 50% de l'énergie maximum de l'énoncé. Lorsque l'énoncé ne contient pas au moins 2 « voyelles accentuées » qui remplissent ces critères, les segments qui correspondent aux 60% de la durée au centre des 3 syllabes les plus longues sont considérés comme « voyelles accentuées ».

Tableau 1. Etapes de la segmentation automatique

La deuxième étape de l'analyse consiste à calculer les moyennes à long terme de différentes caractéristiques vocales à l'intérieur d'unités phonétiques spécifiques. Les différents paramètres et leurs descriptions sont présentés dans le **tableau 2**.

[5] Pour une détection fiable des voyelles, il serait en principe nécessaire d'utiliser un système de reconnaissance de la parole. Mais ces systèmes ne sont pas encore entièrement opérationnels, en particulier pour le discours émotionnel.

Domaine acoustique	**Mesures**	**Segments (sur lesquels les mesures sont réalisées)**
F0	Moyenne, écart-type, maximum, minimum	la totalité d'un énoncé
Durées	Durée (en secondes). Des rapports entre la durée des segments de parole et des silences et entre la durée des parties voisées et non-voisées sont calculés.	la totalité d'un énoncé les pauses (silences) les parties voisées les parties non-voisées les « voyelles accentuées »
Distribution de l'énergie dans le spectre	Moyennes et écarts-types de l'énergie pour l'ensemble du spectre et pour différentes bandes de fréquence. La proportion d'énergie contenue dans les différentes bandes est calculée.[6]	la totalité d'un énoncé les parties voisées les parties non-voisées les « voyelles accentuées »
Proportion d'énergie voisée / non-voisée	Le rapport entre l'énergie moyenne dans les parties voisées et non-voisées pour différentes bandes de fréquences.[7]	les parties voisées et non-voisées des énoncés
Rapport entre énergie dans les hautes et basses fréquences du spectre	proportion d'énergie en dessous de 500 Hz différence entre l'énergie max. de 0 à 2kHz et l'énergie max. de 2 à 5 kHz (index Hammarberg) pente spectrale (au-dessus de 1 kHz)	les parties voisées des énoncés
Distribution du bruit dans le signal voisé (Hilbert Envelope)	Filtrage inverse, transformation de Fourrier sur le signal résiduel, filtrage en 8 bandes (Hamming), transformation de Fourrier inverse, corrélations du 1er signal résultant avec chacun des 7 signaux suivants.[8]	« voyelles accentuées » (soutenues)

Tableau 2. Paramètres acoustiques mesurés à long terme pour différents segments

[6] Les valeurs d'énergie absolue dépendent de l'amplification réalisée lors de l'enregistrement et de la distance entre le locuteur et le microphone. En conséquence, pour chaque segment, seuls les rapports (%) entre la moyenne de l'énergie contenue dans chaque bande de fréquence et la moyenne de l'énergie dans la totalité du spectre sont utilisés. Le spectre peut être fractionné en différentes bandes, par défaut des bandes de 1kHz sont utilisées.

[7] Dans les régions de moyennes et hautes fréquences, ce rapport représente un indicateur de l'intensité vocale. Il y a relativement peu d'énergie voisée dans les hautes fréquences lorsque la voix est faible (soft), alors que pour les voix fortes (loud), l'énergie voisée dans les hautes fréquences augmente relativement à l'énergie non-voisée.

[8] Il est peu probable que du bruit apparaisse dans les 2 premiers signaux. L'apparition de bruit dans les signaux supérieurs se traduira par une diminution de la corrélation.

Dans la littérature aussi bien que dans notre illustration, l'analyse acoustique se limite en général à l'extraction de paramètres segmentaux, souvent agrégés pour la totalité d'un énoncé ou même pour des sections de parole plus importantes. Pourtant, bien que ces paramètres dépassent les limites des segments et peuvent de ce fait indiquer des caractéristiques stables de la phonation et de l'articulation, ils ne font guère ressortir les caractéristiques suprasegmentales, particulièrement la prosodie, de la parole. Or, différents auteurs affirment que les aspects prosodiques de la parole dont l'intonation, l'accentuation et le rythme jouent un rôle prépondérant dans la communication vocale des émotions (v. Fónagy 1983 ; Martin 1987).

4. Le rôle de l'intonation dans l'encodage émotionnel

L'utilisation de moyennes à long terme de paramètres acoustiques segmentaux contribue probablement à 'diluer' l'effet de la réaction émotionnelle sur la voix. Selon le modèle d'encodage proposé par Scherer (1986), les expressions vocales devraient se modifier très rapidement au fil de l'évolution de l'état émotionnel du locuteur. Les changements cognitifs, physiologiques et somatiques qui interviennent au cours d'une réaction émotionnelle sont extrêmement rapides et affectent théoriquement les caractéristiques vocales de manière séquentielle et continue (v. Scherer 1986). En outre, ce modèle ne considère que les changements vocaux qui découlent directement des modifications périphériques associées aux réactions émotionnelles. Or la communication émotionnelle inclut également des modifications vocales destinées à produire un effet sur les interlocuteurs. Cette composante des expressions vocales émotionnelles peut être, dans certains cas, contrôlée volontairement par les locuteurs qui surimposent probablement dans ce cas des modifications stéréotypiques à leurs expressions afin d'influencer leurs interlocuteurs. Il est également possible que cette composante des expressions se traduise parfois par des modifications vocales qui infléchiraient l'intonation de la parole à la manière des actes de parole, c'est-à-dire en modifiant le contour intonatif d'un énoncé. Pour parvenir à une meilleure description des caractéristiques vocales qui expriment l'émotion dans la voix, il est donc très important de développer l'utilisation non seulement de mesures qui rendent compte des modifications du timbre vocal mais également de mesures qui reflètent l'intonation des expressions.

4.1. Définition de l'intonation

Le terme *intonation* est utilisé dans de nombreux contextes et recouvre différentes significations. Dans une définition publiée en 1970, Léon et

Martin évoquent déjà les usages multiples de ce terme. 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 n'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 et Martin 1970 : XV). Comme beaucoup d'autres auteurs, Léon et Martin commencent par définir l'intonation comme recouvrant les variations de durée, de hauteur, et d'intensité. Ils relèvent que ces 3 paramètres ne sont pas indépendants, mais ils restreignent aussitôt l'usage du terme dans leur cadre aux variations de hauteur uniquement. Cet usage plus restreint du terme *intonation* reste le plus fréquent dans les études linguistiques de l'intonation, la plupart du temps en effet les études de l'*intonation* s'attachent essentiellement ou uniquement à la description des fluctuations de la hauteur perçue de la parole.

Léon et Martin mentionnent également que les termes intonation, hauteur, intensité et durée peuvent être utilisés aussi bien dans le domaine de la production vocale que dans le domaine de la perception vocale. Dans la tradition linguistique, très peu d'attention est accordée aux différences qui ne sont pas perçues. En outre, dans le domaine des variations perceptibles, l'intérêt est plutôt dirigé vers les différences qui portent une signification. Dès lors, une différence perçue entre deux contours d'intonation[9] appliqués à un même énoncé ne sera prise en compte que si elle confère une signification (sémantique ou pragmatique) différente à l'énoncé. La plupart des études systématiques et des modèles de l'intonation étant issues de ce courant de recherche, l'intonation (l'évolution de la hauteur perçue) est donc étudiée essentiellement d'un point de vue perceptif et distinctif.

[9] Le contour d'intonation est défini ici comme l'évolution de la hauteur perçue au cours d'un énoncé.

4.2. Les modèles linguistiques de l'intonation

Léon et Martin distinguent déjà en 1970 de nombreux systèmes de transcription de la hauteur perçue. 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.

On peut regrouper aujourd'hui les modèles linguistiques de l'intonation en deux catégories principales. La première catégorie recouvre les modèles qui conçoivent l'intonation comme une succession de tons (ou de courbes) distinctifs sur le plan phonologique. Il s'agit de modèles qui sont plus ou moins directement issus des systèmes de transcription décrits ci-dessus. La deuxième catégorie recouvre les modèles qui conçoivent le contour intonatif comme le résultat de la superposition de plusieurs composantes. Le modèle type de cette catégorie est le modèle à deux composantes (Ladd 1983), où la première composante correspond à une « ligne de déclinaison »[10] (une droite qui en théorie décline lentement du début à la fin d'une production vocale) et la deuxième composante aux variations de hauteur à plus court terme (les accents locaux qui se superposent à la ligne de déclinaison).

4.3. Les limites des modèles linguistiques pour l'étude de l'intonation émotionnelle

La difficulté principale dans le domaine de l'étude de l'encodage de l'intonation émotionnelle consiste à définir un « système de codage » de l'intonation qui permette ensuite d'évaluer l'effet de différentes émotions sur l'intonation. Les transcriptions de la hauteur perçue en séquences de tons ou de courbes fournissent des descriptions en configurations qui

[10] Voir Grobet & Simon (ici même).

pourraient être utilisées à cette fin. Ces transcriptions présentent toutefois un certain nombre d'inconvénients. En premier lieu, elles sont entièrement orientées vers la description de la perception de l'intonation et non de la production de l'intonation. Les transcriptions de la hauteur perçue en séquences de tons (ou de courbes) intègrent de manière implicite les paramètres de durée et d'intensité. La hauteur perçue est en effet dépendante non seulement de la fréquence fondamentale mais également de la durée et de l'amplitude d'un segment de parole. L'un des inconvénients des ces transcriptions réside donc dans l'absence de spécification des caractéristiques physiques (de fréquence fondamentale, de durée et d'amplitude) qui sont à l'origine des contours de hauteur perçue. D'autre part, afin de caractériser les effets de différentes émotions sur l'intonation un système de description quantifiable est idéalement préférable à un système produisant des descriptions catégorielles. Lors de la transcription des contours de hauteur perçue, les tons ou les courbes sont en général notés de manière relative à l'expression décrite et non en référence à une valeur absolue. Or – pour évaluer l'effet de différentes émotions sur l'intonation – on souhaiterait pouvoir comparer plusieurs expressions non seulement en ce qui concerne la forme globale de leurs contours mais également en ce qui concerne par exemple leurs hauteurs de référence respectives ou encore leurs hauteurs finales respectives.

4.4. Perspectives méthodologiques pour les recherches actuelles et futures

Afin de décrire l'intonation d'un grand nombre expressions et d'identifier les caractéristiques de l'intonation liées à l'expressions de différentes émotions, il est nécessaire de développer des systèmes de mesures permettant de décrire des corpus importants. Afin de permettre un traitement statistique des résultats, les mesures devront être, de préférence, quantifiables. Le « codage » des contours intonatifs peut être réalisé à « la main » ou automatisé.

4.4.1 Stylisations automatiques des contours

Un ensemble d'analyses automatiques des contours d'intensité et des contours de fréquence fondamentale sont présentées dans le **tableau 3**. Les programmes utilisés pour l'extraction automatique des paramètres décrits ont été développés dans notre groupe de recherche par G. Klasmeyer (pour plus de détails v. Klasmeyer 2000).

	Etapes de la stylisation	**Segments**	**Mesures**
Contours de F0	Le contour continu (les parties non-voisées sont remplacées par des droites) de F0 est filtré pour éliminer les micro-fluctuations. Les segments non-voisés sont ensuite ramenés à zéro. Des lignes droites sont tracées entre les minima et les maxima locaux considérés comme « importants »[11]..	- montées - parties plates hautes (au-dessus de la moyenne locale) - descentes - parties plates basses (au-dessous de la moyenne locale) - parties non-voisées	- nombre de segments de chaque type - moyenne et écart-type de leurs durées - moyenne et écart-type de leurs pentes (excepté pour les parties non-voisées) --- - position du maximum absolu dans l'énoncé
Contours d'énergie	Le contour d'énergie est filtré à 30 Hz pour éliminer les micro-fluctuations. Les valeurs inférieures à 1% du maximum d'énergie dans l'énoncé sont ramenées à zéro (pauses). Des lignes droites sont tracées entre les minima et les maxima locaux considérés comme « importants »	- montées (*onsets*) - parties plates hautes (au-dessus de la moyenne locale) - descentes (*decays*) - parties plates basses (au-dessous de la moyenne locale) - pauses	- nombre de segments de chaque type - moyenne et écart-type de leurs durées - moyenne et écart-type de leurs pentes (excepté pour les pauses) --- - position du maximum absolu dans l'énoncé

Tableau 3. Variables issues de la stylisation automatique des contours

Une approche alternative à la stylisation automatique consiste à réaliser une stylisation manuelle du contour de F0 au travers de l'identification de certains points jugés importants a priori. Cette approche peut être préférée à la stylisation automatique lorsque les corpus ne sont pas trop grands, elle permet de corriger les erreurs inhérentes à la détection automatique de la fréquence fondamentale (par exemple la détection de périodicité sur des segments non-voisés) et de prendre en compte les segments phonétiques sur lesquels les excursions de F0 sont réalisées. A titre d'exemple, nous décrivons ci-dessous des critères de codage utilisés par l'un des auteurs de cet article (T. Bänziger) pour le codage de la F0 d'un ensemble d'expressions émotionnelles.

[11] « L'importance » des maxima et minima locaux dépend de la forme globale du contour, les critères de décision sont décrits dans Klasmeyer 2000.

4.4.2 Stylisation manuelle du contour de la F0

Les points retenus pour la description des contours de F0 correspondent aux valeurs en Hz (absolues) et en secondes (relatives au début de chaque expression) relevées pour le commencement et la fin du contour, pour un premier « accent », pour un deuxième « accent » et pour un « accent » final (le plus souvent une descente ou parfois une montée).

La stylisation manuelle du contour de la fréquence fondamentale a été effectué pour 144 enregistrements. Ces enregistrements ont été extraits d'une base de données constituée et décrite en détail par Banse et Scherer (1996). Des enregistrements produits par 9 acteurs ont été sélectionnés. Tous les acteurs prononcent deux séquences de 7 syllabes sans signification[12] et expriment 8 types d'émotions : colère chaude et colère froide, anxiété et peur panique, tristesse et désespoir, joie calme et joie intense.

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 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, troisième et quatrième points 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 « gött laich ». Les points cinq, six et sept 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 « accent », le premier minimum correspond au point où la pente de la fréquence fondamentale devient positive. Au cas

[12] 1. « hät san dig prong nju ven tsi », 2. « fi gött laich jean kill gos terr »

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, neuf et dix ; correspondent à « l'accent final » de chaque séquence ; les minimum, maximum, minimum locaux sont relevés pour les syllabes « tsi » et « ter ». Sur ces syllabes uniques, il est très 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 », « götte 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 considéré, les points 4, 7 et 10 sont considérés comme données manquantes.

Ce codage de la fréquence fondamentale permet de définir le type d'excursion (montée suivie d'une descente, descente ou montée uniquement, absence d'excursion) réalisé sur un segment de parole prédéfini ; dans notre exemple, il s'agit 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. Contrairement aux contours décrits en configurations de tons, ce codage permet de réaliser un traitement statistique des caractéristiques des contours.

A travers les exemples qui précèdent, nous avons voulu souligner la nécessité de développer des mesures qui permettent de parvenir à une meilleure description des expressions émotionnelles et, en conséquence, à une meilleure différentiation des émotions exprimées. Les pages qui suivent sont consacrées à la perspective du décodage de l'émotion exprimée. Dans ce domaine également, il apparaît nécessaire d'accorder

plus d'attention aux caractéristiques vocales qui sont perçues comme émotionnelles.

5. Etude du décodage de l'émotion dans la voix

De nombreuses recherches ont été consacrées à l'étude de la reconnaissance des émotions communiquées par la voix. Ces recherches ont montré que les émotions – souvent exprimées par des acteurs – sont très bien identifiées par des auditeurs chargés de choisir parmi un ensemble de termes émotionnels celui qui décrit le mieux l'émotion exprimée. Dans ces études, des pourcentages de reconnaissance correcte sont calculés pour chaque émotion considérée. Une revue de la littérature basée sur environ 30 études réalisées avant le milieu des années 1980 (Scherer 1989) indique que le pourcentage de reconnaissance correcte moyen est d'environ 60%, soit environ 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, Banse et Wallbott (2001) rapportent un pourcentage moyen de reconnaissance correcte de 62% pour 11 études effectuées dans différents pays occidentaux. Les taux de reconnaissance correcte varient en fonction de l'émotion exprimée et sont remarquablement constants d'une étude à l'autre. Les expressions de tristesse et de colère sont en général mieux reconnues que les expressions de peur ou de joie. Lorsque le dégoût fait partie des émotions étudiées, les expressions correspondantes sont toujours moins bien reconnues que celles correspondant aux autres émotions étudiées.

Relativement au grand nombre d'études portant sur la reconnaissance des émotions dans les expressions vocales, peu d'études se sont intéressées aux *processus* de décodage. Il n'existe donc pas beaucoup de données concernant les caractéristiques vocales utilisées par les auditeurs lors du décodage de l'émotion. Les études qui ont été effectuées dans ce domaine ont utilisé trois sortes d'approches. Quelques études ont établi des corrélations multiples entre les caractéristiques acoustiques des expressions 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. 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 via la synthèse ou la re-synthèse. Cette dernière approche est à l'heure actuelle la plus prometteuse. Elle 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. 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. Les paragraphes suivants sont en conséquence consacrés à la présentation de quelques exemples issus de la littérature pertinente.

5.1. Le rôle de l'intonation dans le décodage

Différents auteurs ont émis des propositions concernant les caractéristiques des contours intonatifs qui seraient perçus comme émotionnels. Dans ce domaine, il existe deux types d'approches sensiblement différentes. Certains auteurs défendent l'existence d'associations entre des formes intonatives et des significations émotionnelles en se basant sur des exemples qu'ils transcrivent puis soumettent à l'appréciation de leurs lecteurs. D'autres auteurs basent leurs descriptions sur le traitement statistique des attributions émotionnelles effectuées par des groupes d'auditeurs relativement à des expressions réalisées par différents locuteurs ou souvent manipulées par les auteurs.

Les descriptions de Fónagy et Magdics (1963) donnent une illustration du premier type d'approche. Ces auteurs décrivent l'évolution de la hauteur perçue par une succession de tons sur une portée musicale pour différents énoncés (exemples) qui correspondent à différentes situations émotionnelles. Frick (1985) a émis une série de critiques relativement à ce type d'approche. Il cite dans cette catégorie les travaux de Halliday (1970), Jassem (1952), O'Connor & Arnold (1973) et Schubiger (1958). Selon

Frick, ces auteurs attribuent une signification spécifique à un contour intonatif particulier. Ils fournissent ensuite un exemple contextualisé de l'utilisation de ce contour qui a pour fonction de confirmer que le contour possède effectivement la signification postulée. L'utilisation d'un exemple afin de démontrer qu'un contour spécifique possède une certaine signification pose quelques problèmes. Le contenu verbal contient souvent la signification que le contour devrait en principe transmettre[13]; et 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.

Généralement, les études de l'intonation émotionnelles fondées sur le deuxième type d'approche ne tentent pas de faire correspondre un contour ou plusieurs contours intonatifs à une ou plusieurs significations. Ces études essaient plutôt d'établir des liens statistiques (c'est-à-dire probabilistes et non systématiques) entre différentes caractéristiques de l'intonation de différents énoncé et l'attribution émotionnelle.

Une étude classique dans ce domaine a été publiée par Lieberman et Michaels en 1962. Ces auteurs ont utilisé des expressions correspondant à 8 « modes émotionnels »[14] produits pour 8 énoncés anglais différents. 85% des modes sont reconnus correctement par un groupe d'auditeurs lorsque les enregistrements originaux leur sont présentés. Les auteurs ont re-synthétisé la fréquence fondamentale de ces expressions sur une voyelle fixe, cette opération conserve toute l'information relative à l'évolution dans le temps de la fréquence fondamentale mais supprime totalement les modifications spectrales et les variations d'amplitudes des expressions originales. Lorsque seule l'information relative au contour de F0 est encore présente, la reconnaissance correcte globale des différents modes est encore de 44%. En ajoutant à cette information les variations d'amplitude le pourcentage de reconnaissance correcte n'augmente que de 3% (à 47% de reconnaissance correcte). En revanche, le taux global de reconnaissance correcte diminue sensiblement lorsque les variations à très court terme du contour de F0 sont supprimées. Le « lissage » du contour de F0 fait chuter la reconnaissance correcte à 25%. La re-synthèse du volume uniquement permet encore une identification correcte de 14% des expressions. Cette

---

[13] Chez Fónagy et Magdics, par exemple, l'énoncé « Comme je suis heureuse de te voir! Je ne pensais pas te rencontrer! » est utilisé pour illustrer un contour typique de joie.

[14] 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'

étude indique que le contour de F0, le contour d'amplitude et les informations spectrales contribuent à la reconnaissance des caractéristiques non-verbales exprimées. Elle souligne également que les variations fines du contour de fréquence fondamentale jouent un rôle important dans ce domaine.

Les résultats de Lieberman et Michaels ne donnent pas d'indication concernant les caractéristiques de la courbe de F0 qui sont utilisées par les auditeurs pour identifier spécifiquement l'un ou l'autre « mode émotionnel ». Uldall (1964) a publié une autre étude pionnière dans laquelle 16 contours de F0 ont été appliqués 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, Ladd et Silverman (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 (par exemple le type de question, « yes/no » versus « WH »). 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 entre par exemple en corrélation avec les jugements du degré d'activation émotionnelle des locuteurs. Les auteurs relèvent que les variables qui affectent les attributions émotionnelles de manière continue reflètent 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 tendent à signaler plutôt des attitudes du locuteur (par exemple amicale ou réprobatrice).

5.2. Perspectives méthodologiques

Les quelques travaux présentés ci-dessus illustrent la possibilité d'influencer les attributions émotionnelles en manipulant la courbe de F0 (v. aussi Scherer & Oshinsky 1977; Ladd et al. 1985; Mozziconacci 1998). Par ailleurs, les techniques de synthèse de la parole permettent de générer des stimuli très contrôlés qui peuvent être utilisés dans le cadre d'études neuropsychologiques afin d'identifier les structures cérébrales et les processus corticaux et sous-corticaux qui sous-tendent la perception de l'intonation. Une illustration de l'utilisation des techniques de synthèse dans ce domaine est présentée ci-dessous.

Des travaux récents en neuropsychologie clinique ont permis de distinguer, au niveau du système nerveux central, le traitement des aspects linguistiques segmentaux, du traitement des aspects linguistiques suprasegmentaux (Ross 1981 ; Pell 1998 ; van Lanker & Sidtis 1992 ; Heilman, Bowers, Speedie & Coslett 1984). L'intonation peut avoir une fonction pragmatique (linguistique) lorsqu'elle permet, par exemple, de distinguer un énoncé interrogatif d'un énoncé affirmatif, mais elle peut également, comme nous l'avons observé dans ce chapitre, transmettre une information de type émotionnelle. Une voie de recherche actuelle, générant de nombreuses études, concerne la spécialisation hémisphérique de la perception de l'intonation linguistique (pragmatique) et de l'intonation émotionnelle. Une étude en cours, réalisée en collaboration entre la faculté de psychologie de l'Université de Genève (D. Grandjean et K. Scherer) et le département de Neurologie et de Neuropsychologie de l'Hôpital Cantonal de Genève (C. Ducommun, C. Michel et T. Landis), explore ces différentes dimensions de la perception de l'intonation à travers les techniques d'électroencéphalographie et de cartographie cérébrale. Cette étude s'articule autour de la comparaison des patterns électrophysiologiques et des régions cérébrales impliquées dans la genèse de ces patterns pour trois niveaux de discriminations différents : discrimination de l'intonation linguistique ou pragmatique, discrimination

de l'intonation émotionnelle et discrimination phonémique. Dans le cadre de ce chapitre, l'élaboration des stimuli employés dans le paradigme électrophysiologique sera présentée.

Comme mentionné ci-dessus, la synthèse vocale permet une modulation très fine et contrôlée des paramètres acoustiques permettant de caractériser différents types d'intonations, cette technique a été préférée aux enregistrements d'acteurs, introduisant un grand nombre de modifications des paramètres acoustiques non contrôlés. La synthèse vocale a été effectuée grâce au logiciel Mbrola (Dutoit 1997) élaboré par la faculté polytechnique de Mons en Belgique. Pour des raisons pratiques, liées à l'électroencéphalographie, et pour minimiser les effets linguistiques, seuls trois mots phonétiquement proches : « vallon », « talon » et « ballon » ont été synthétisés. La synthèse initiale, sans accent d'intonation particulier, a conduit à la production de ces trois mots sur une durée de 400 ms pour chaque mot ; 100 ms par phonème. Cette durée initiale des stimuli a été réduite à une durée de 350 ms en pratiquant une compression temporelle. Sur la base des prédictions de Banse & Scherer (1996), deux paramètres ont été modifiés sur ces stimuli initiaux : la hauteur (F0) et l'enveloppe des stimuli « neutre » afin de produire les stimuli expérimentaux. Le **tableau 4** présente les caractéristiques des stimuli synthétisées avec Mbrola pour les deux conditions de prosodie linguistique et émotionnelle.

Mot		/ *balo~* /			
Phonèmes		[b]	[a]	[l]	[o~]
	Neutre	100 ms 100% = 130 Hz	100 ms 100% = 130 Hz	100 ms 100% = 130 Hz	100 ms 100% = 130 Hz
Prosodie linguistique	Interrogatif	75 ms 100% = 130 Hz	75 ms 100% = 130 Hz	50 ms 100% = 130 Hz	200 ms 25% = 140 Hz 50% = 200 Hz 75% = 260 Hz
Prosodie linguistique	Affirmatif	75 ms 100% = 130 Hz	75 ms 25% = 175 Hz 50% = 175 Hz 75% = 175 HZ	50 ms non spécifié	200 ms 25% = 120 Hz 50% = 80 Hz 75% = 50 Hz
Prosodie émotionnelle	Joie	75 ms 100% = 130 Hz	75 ms 25% = 140 Hz 50 % = 200 Hz 75 % = 260 Hz	50 ms non spécifié	200 ms 75 % = 130 Hz
Prosodie émotionnelle	Triste	75 ms 100% = 130 Hz	150 ms 5% = 70 Hz 25% = 65 Hz	25 ms non spécifié	150 ms 10% = 60 Hz 50% = 60 Hz 90% = 60 Hz

Les % indiquent la durée relative pour laquelle la F0 a été changée pour certains phonèmes. Lorsque nous n'avons pas modifié les paramètres pour un phonème la mention « non spécifié » le précise.

Tableau 4. Caractéristiques des stimuli de synthèse
(durées et modulation de la F0 des phonèmes)

Globalement le contour de F0 augmente pour les énoncés de type « joie » alors qu'il descend pour les énoncés « triste ». Il augmente sur le dernier phonème de l'énoncé pour les types « interrogatif ». Pour les énoncés du type « affirmatif » le contour monte sur le deuxième phonème pour redescendre sur le dernier phonème

Les **figures 1 et 2** présentent les différents tracés de fréquence fondamentale pour les différentes conditions « neutre », « joie » et « triste » d'une part, et « interrogatif », « affirmatif » et « neutre » d'autre part, sur le mot « ballon ».

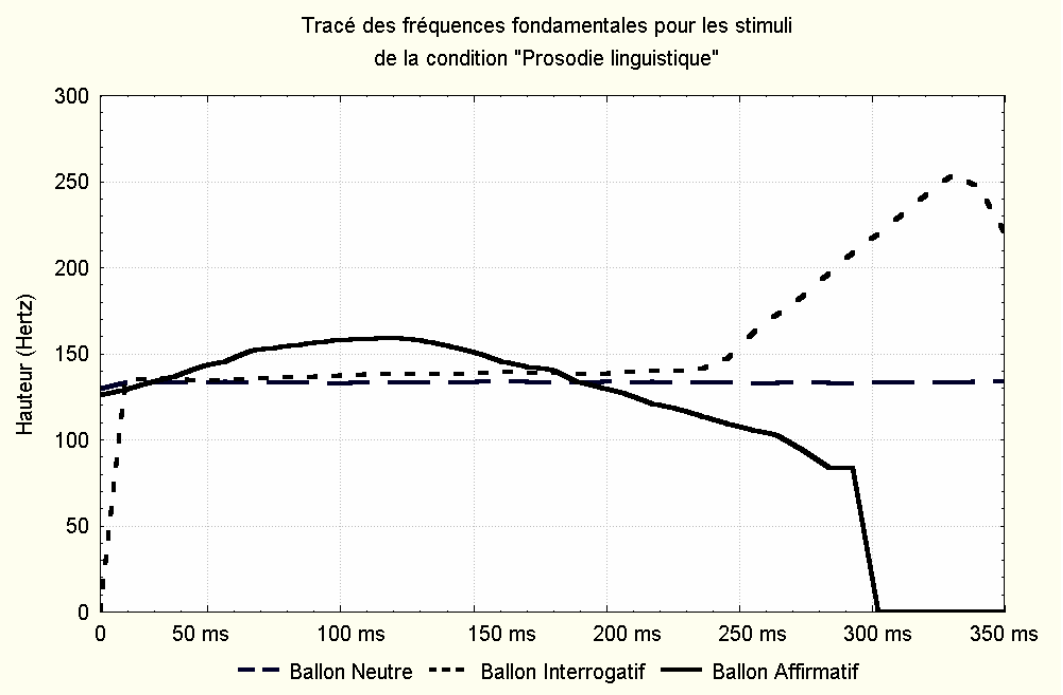

Figure 1

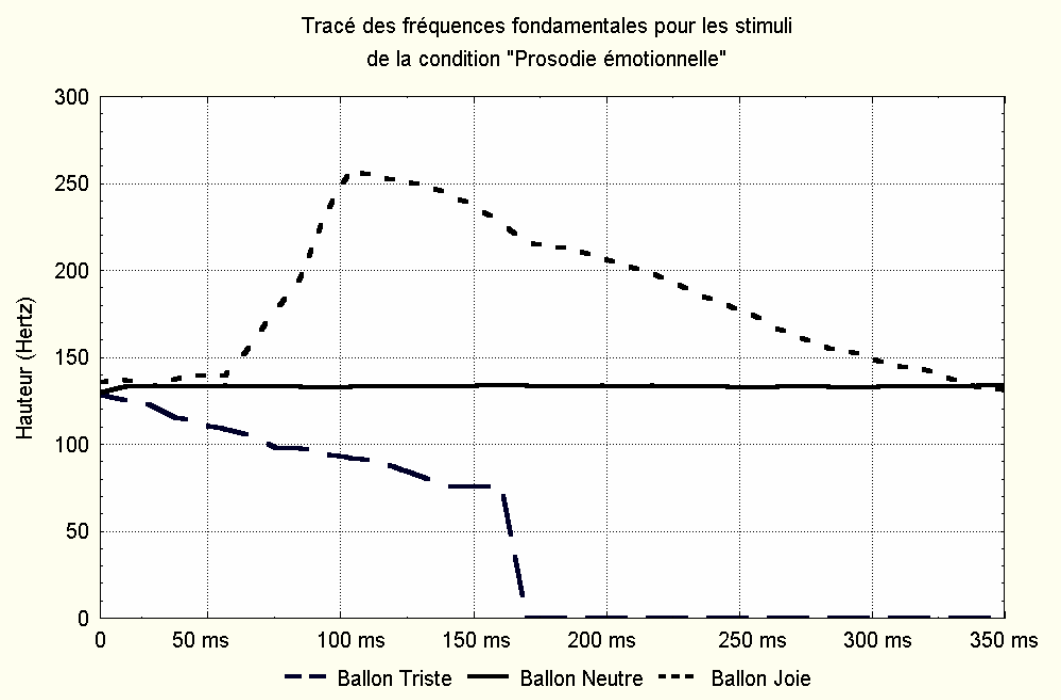

Figure 2

Les trois mots utilisés ont été modifiés de manière similaire au mot « ballon » présenté ci-dessus, toutefois quelques différences mineures

existent liées aux premiers phonèmes qui sont différents pour les trois mots.

Afin de s'assurer d'une bonne discrimination des différents énoncés dans les différentes conditions prosodiques, une étude de jugement a été réalisée visant à valider les stimuli pour leurs utilisations dans le paradigme électroencéphalographique. Vingt sujets devaient, dans une première phase, évaluer sur une échelle continue le degré d'émotion contenu dans les énoncés avec prosodie émotionnelle. Dans une deuxième phase ils devaient identifier les différents niveaux de prosodie linguistique puis les différents niveaux de prosodie émotionnelle de manière catégorielle. L'évaluation du degré d'émotion dans les énoncés est réalisée sur cinq échelles : joie, triste, content, ennui et colère. Comme le montre la **figure 3**, les stimuli « joie » obtiennent un score significativement plus élevé sur les émotions « joie » et « contentement » que sur toutes les autres émotions ($F(4,452) = 488.59$, $p<.01$). Les stimuli « triste » obtiennent un score élevé sur les dimensions « tristesse » et « ennui » et très bas sur les émotions positives « contentement » et « joie » ($F(4,452)=123.02$, $p>.01$), notez que ces stimuli sont également associés à l'émotion « colère ». Les stimuli « neutre » obtiennent également des moyennes significativement différentes selon les différentes émotions, les plus élevées pour les émotions de « tristesse » et « ennui ».

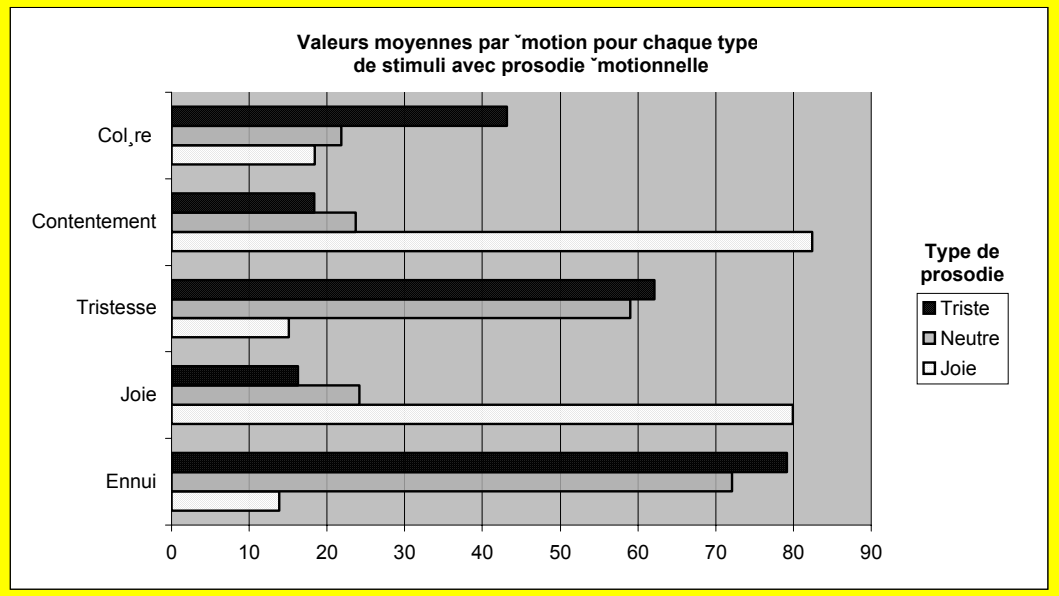

Figure 3

La tâche en électroencéphalographie consistant en une identification entre trois types de prosodie linguistique et émotionnelle, les stimuli ont été également pré-testés dans le cadre de ce paradigme. Dans cette deuxième

phase, la tâche des sujets était d'identifier la prosodie d'un stimulus avec trois types de prosodie à choix. Pour la prosodie émotionnelle : joie, triste ou neutre et pour la prosodie linguistique : affirmatif, interrogatif ou neutre. Une tâche d'identification phonémique a été également réalisée. Les résultats présentés dans la **figure 4** montrent une très bonne identification des différentes prosodies. Dans la tâche d'identification phonémique, non représentée dans la figure 4, le taux de reconnaissance correcte est de 99.5% pour les trois mots.

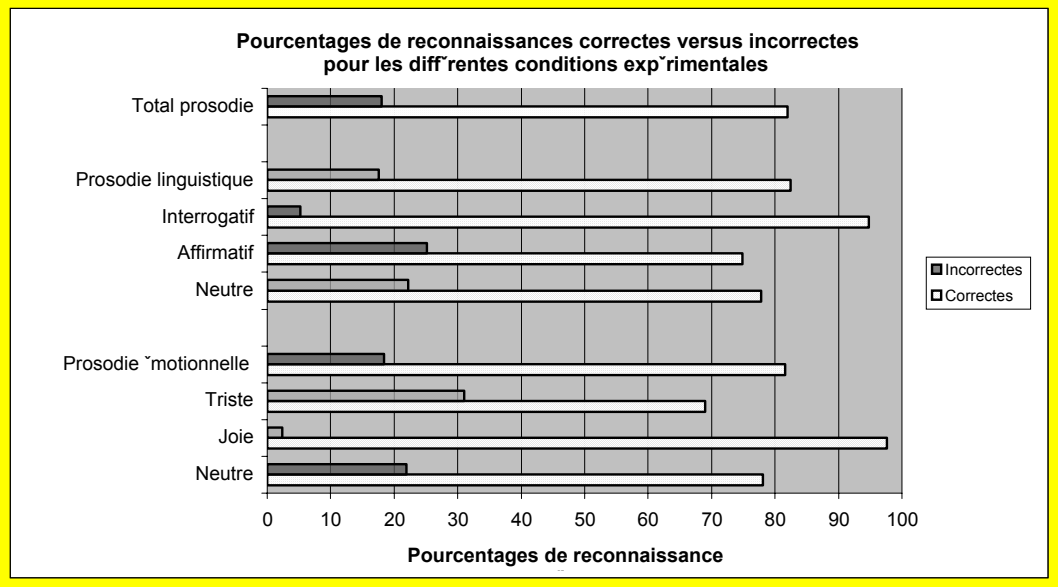

Figure 4

Les résultats ci-dessus indiquent clairement qu'il est tout à fait possible, à partir de modifications de quelques paramètres acoustiques relativement simples, de simuler différentes prosodies, qui sont suffisamment bien identifiées pour être utilisées dans d'autres domaines, comme l'imagerie cérébrale par exemple. L'utilisation de telles techniques – permettant d'une part de contrôler précisément les différences acoustiques entre les stimuli et d'autre part de mettre en relation ces modifications acoustiques avec des jugements effectués par des sujets – pourrait donner lieu à de nouvelles perspectives. Ainsi, il serait possible de modifier progressivement les différents paramètres acoustiques pertinents pour la discrimination de différentes prosodies émotionnelles et déterminer, pour certaines émotions, où se situent le ou les points de différenciation. Il est probable que certains paramètres sont plus pertinents que d'autres pour des émotions spécifiques. Les pondérations de modifications pourraient également être différentes pour des émotions distinctes ; certaines modifications mineures sur un paramètre acoustique affecteront probablement la perception d'une émotion donnée de manière importante

alors que d'autres modifications plus importantes sur d'autres paramètres n'auront que peu d'effets sur la perception de cette émotion ; et l'inverse pourrait être vrai pour une autre émotion. La modification progressive des paramètres acoustiques pourrait également être utilisée dans l'étude des réseaux neuronaux fonctionnels du système nerveux central impliqués dans la détection des changements acoustiques et dans la catégorisation consécutive. L'utilisation des techniques de synthèse vocale pour l'étude de l'intonation constituera donc assurément une voie d'étude florissante dans les années à venir.

6. Conclusion

Les différents champs de recherche présentés brièvement ci-dessus convergent dans un intérêt renaissant pour la prosodie appréhendée dans une perspective pluridisciplinaire. Les nouvelles méthodes à disposition de la recherche – telles que la synthèse de parole et les techniques d'analyses automatisées – augurent d'un accroissement des études futures et d'une amélioration des connaissances dans ce domaine. Toutefois, un accord préliminaire, concernant aussi bien la définition du phénomène, des modèles théoriques sous-jacents et des variables centrales est indispensable pour une établir une base commune aux différents courants de recherche investiguant ce domaine d'étude. Dans le contexte de la recherche sur la communication vocale des émotions, il semble particulièrement important d'insérer l'étude de l'intonation – qui a été longtemps appréhendée par des exemples isolés ou des illustrations fabriquées – dans une perspective de recherche empirique. A cette fin, il reste nécessaire de développer à la fois des modèles et des hypothèses, ainsi que des mesures quantitatives de l'intonation permettant de soumettre les prédictions découlant des modèles à des tests expérimentaux.

Bibliographie

BANSE R. & SCHERER K. (1996), « Acoustic profiles in vocal emotion expression », *Journal of Personality and Social Psychology* 70, 614-636.

BOERSMA P. & WEENINK D.J.M. (1996), Praat, a system for doing phonetics by computer, version 3.4, Institute of Phonetic Sciences of the University of Amsterdam, Report 132.

DUTOIT T. (1997), *An Introduction to Text-To-Speech Synthesis*, Dordrecht, Kluwer Academic Publishers.

FONAGY I. (1983), *La vive voix*, Paris, Payot.

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, 412-429.

HALLIDAY M.A.K. (1970), *A course in spoken English: Intonation*, London, Oxford University Press.

HEILMAN K.M., BOWERS D., SPEEDIE L. & COSLETT H.B. (1984), « Comprehension of affective and non affective prosody », *Neurology* 34, 917-921.

JASSEM W. (1952), *Intonation of conversational English (educated Southern British)*, Wroclaw, Travaux de la société des sciences et des lettres de Wroclaw.

JOHNSTONE T. & SCHERER K.R. (2000), « Vocal communication of emotion », in Lewis M. & Haviland-Jones J. (éds), *Handbook of Emotions*, *Second Edition*, New York, Guilford Press, 220-235.

KLASMEYER G. (1999), « Akustische Korrelate des stimmlich emotionalen Ausdrucks in der Lautsprache », Wodarz H.-W., Heike G., Janota P. & Mangold M. (éds), *Forum Phoneticum* 67, Frankfurt am Main, Hector.

KLASMEYER G. (2000), « An automatic description tool for time-contours and long-term average voice features in large emotional speech databases », *Proceedings of ISCA workshop on Speech and Emotion*, Newcastle

LADD D.R. (1983), « Peak features and Overall Slope », in Cutler A. & Ladd D.R. (éds), *Prosody: models and measurments*, Berlin, Springer-Verlag, 39-52.

LADD D.R., SILVERMAN K.E.A., TOLKMITT F., BERGMANN G. & SCHERER K.R. (1985), « Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect », *Journal of the Acoustical Society of America* 78, 435-444.

LAVER J. (1980), *The phonetic description of voice quality*, Cambridge, Cambridge University Press.

LÉON P.R. & MARTIN PH. (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, 922-927.

MARTIN P. (1987), « Prosodic and rhythmic structures in french », *Linguistics* 25, 925-949.

MOZZICONACCI S.J.L. (1998), *Speech variability and emotion: production and perception*, Netherlands, University of Eindhoven.

O'CONNOR J.D. & ARNOLD G.F. (1973), *Intonation of colloquial English* (2nd ed.), London, Longman.

PELL M.D. (1998), « Recognition of prosody following unilateral brain lesion: influence of functional and structural attributes of prosodic contours », *Neuropsychologia* 36, 701-715.

ROSS E.D. (1981), « The aprosodias: functional-anatomic organization of the affective components of language in the right hemisphere », *Archives of Neurology* 38, 561-569.

SCHERER K.R. (1986), « Vocal Affect Expression: A Review and a Model for Future Research », *Pychological Bulletin* 99, 143-165.

SCHERER K.R. (1989), « Vocal correlates of emotion », in Wagner H. & Manstead A. (éds), *Handbook of psychophysiology: Emotion and social behavior*, London, Wiley, 165-197.

SCHERER K.R., BANSE R. & WALLBOTT H.G. (2001), « Emotion inferences from vocal expression correlate across languages and cultures », *Journal of Cross-Cultural Psychology* 32, 76-92.

SCHERER K.R., FELDSTEIN S., BOND R.N. & ROSENTHAL R. (1985), « Vocal cues to deception: A comparative channel approach », *Journal of Psycholinguistic Research* 14, 409-425.

SCHERER K.R., JOHNSTONE T. & KLASMEYER G. (2002), « Vocal expression of emotion », in Davidson R.J., Goldsmith H. & Scherer K.R. (éds), *Handbook of the Affective Sciences*, New York, 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.S. (1977), « Cue utilization in emotion attribution from auditory stimuli », *Motivation and Emotion* 1, 331-346.

SCHUBIGER M. (1958), *English Intonation: Its form and function*, Tübingen, Max Niemeyer.

ULDALL E. (1964), « Dimensions of meaning in intonation », in Abercrombie D., Fry D.B., MacCarthy P.A.D., Scott N.C. & Trim J.L.M. (éds), *In honour of Daniel Jones: papers contributed on the Occasion of his Eightieth birthday, 12 september 1961*, London, Longman, 271-279.

VAN LANKER 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, 963-970.