Le rôle de l'intonation dans la communication vocale des émotions : test par la synthèse

Michel Morel, Tania Bänziger

CILL 30 (2004) n° 1-3, pp. 207-232

LE ROLE DE L'INTONATION DANS LA COMMUNICATION VOCALE DES EMOTIONS : TEST PAR LA SYNTHESE

Michel MOREL
CRISCO (CNRS)
Université de Caen

Tanja BÄNZIGER
FAPSE
Université de Genève

1. INTRODUCTION

Il est aujourd'hui bien établi que les émotions et les attitudes peuvent être communiquées avec succès par le biais des expressions vocales non-verbales. Les revues de la littérature dans ce domaine (v. par exemple Scherer 2003, Elfenbein & Ambady 2002) indiquent qu'environ 60% des expressions sont reconnues correctement, bien que cette proportion varie considérablement en fonction du type d'émotion ou d'attitude exprimée, des expressions spécifiques utilisées et des individus testés dans les différentes études réalisées.

Les caractéristiques des expressions vocales qui communiquent les émotions ou les attitudes sont en revanche encore mal connues. Les revues des études qui ont tenté de décrire les expressions vocales correspondant à différentes émotions sur le plan acoustique constatent souvent que les paramètres acoustiques mesurés reflètent l'activation émotionnelle sous-jacente aux réactions émotionnelles exprimées, mais ne semblent pas différencier les émotions correspondant à un même niveau d'activation (Johnstone & Scherer 2000, Scherer 2003). À ce propos, la nécessité de mesurer différents paramètres acoustiques qui reflèteraient des aspects vocaux liés non seulement à l'activation émotionnelle mais également au type d'émotion exprimé a été soulevée à plusieurs reprises (v. par exemple Banse & Scherer 1996, Klasmeyer 1999, Juslin & Laukka 2001). Différents auteurs ont formulé des hypothèses théoriques et des prédictions concernant les aspects vocaux affectés par l'émotion et utilisés pour la communiquer (v. par exemple Scherer 1986, Klasmeyer 1999). Toutefois très peu de données empiriques sont, à ce jour, disponibles relativement à ces prédictions et il n'existe pas de modèle communément accepté relativement aux caractéristiques vocales acous-

tiques ou perçues qui seraient particulièrement importantes pour la communication des émotions.

De multiples indications dans la littérature suggèrent néanmoins que différentes composantes vocales interviennent dans la communication non-verbale des émotions. Les aspects liés à la qualité vocale – composante qui recouvre des caractéristiques liées au mode de phonation (source vocale) et plus généralement au timbre (résonances dans l'appareil de production vocale) – sont généralement distingués des aspects liés à l'intonation – composante qui recouvre les aspects de l'évolution temporelle de la hauteur, de l'intensité et du rythme de la parole. Ces deux composantes ont été aussi parfois distinguées comme appartenant respectivement au domaine segmental (pour la qualité vocale) et au domaine suprasegmental de la parole (pour l'intonation). Cette distinction nous paraît discutable : nous préférons réserver le domaine segmental à la seule voix de base (plate et ne comportant comme variations que la microprosodie – phénomène articulatoire). Nous parlerons alors de *prosodie* pour désigner l'ensemble des variations acoustiques suprasegmentales, à savoir l'intonation et les variations de timbre.

Des tentatives de séparer la contribution respective des deux composantes de la prosodie – intonation et timbre – dans la communication vocale des émotions ont été réalisées en utilisant des méthodes de "dégradation" du signal acoustique destinées à affecter sélectivement le timbre (par exemple par un filtrage "low-pass" du signal acoustique) et l'intonation (en altérant la séquence temporelle du signal acoustique) ou en utilisant des techniques de synthèse vocale afin de modifier sélectivement l'une ou l'autre composante (v. par exemple Scherer, Feldstein, Bond et Rosenthal 1985).

Une étude pionnière dans ce domaine est due à Lieberman et Michaels (1962). Ces auteurs ont tenté de neutraliser la qualité vocale d'un ensemble d'expressions émotionnelles en resynthétisant le contour de F0 et le contour d'intensité de ces expressions sur une voyelle fixe. Ils ont démontré que la proportion de "modes émotionnels" correctement reconnus par des auditeurs est approximativement divisée par deux lorsque le contour de F0 (avec ou sans le contour d'intensité) est isolé de cette manière de la qualité vocale des expressions. Dans le même sens, Ladd, Silverman, Tolkmitt, Bergman & 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. D'autre part, une étude de Scherer, Ladd & Silverman (1984) a montré que parmi un ensemble d'attitudes qui sont reconnues dans des expressions vocales intactes, une seule attitude (la politesse) est encore reconnue lorsque les expressions sont filtrées "low-pass" (à environ 130Hz), alors qu'un nombre plus important d'attitudes et d'émotions (la poli-

tesse, la menace, l'insécurité, une attitude agréable/positive et l'activation) sont préservées dans les expressions dont la séquence temporelle est altérée et la composition spectrale préservée.

Il existe par ailleurs une tradition de recherche rassemblant 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, O'Connor & Arnold 1973, Mozziconacci 1998). L'existence d'interactions très importantes entre les contours intonatifs et le contenu linguistique a été opposée à cette perspective. Une montée finale pourrait, par exemple, être utilisée pour communiquer un doute, mais ne sera interprétée comme telle que si le contenu sémantique et syntaxique s'y prête. Dans ce sens, une étude de Scherer, Ladd & Silverman (1984) a démontré qu'un contour de F0 descendant, lorsque le contexte syntaxique requiert un contour de F0 montant (pour une question totale), recevra une interprétation émotionnelle qui diffère de l'interprétation émotionnelle donnée à une expression avec un contour de F0 descendant lorsqu'un contour montant n'est pas exigée par le contexte syntaxique (pour une question formulée en utilisant un lexème interrogatif de type qui-quand). De même, Uldall (1964) a démontré que la signification qui se dégage de différents contours varie en fonction de la phrase utilisée.

En revanche, ce dernier auteur (Uldall 1964) trouve également que certaines propriétés des contours qu'elle a appliqués sur différentes phrases sont liées aux jugements émotionnels qu'elle a recueillis indépendamment des phrases utilisées. Dans le même sens, une expérience de Papousek (1994) démontre qu'avant l'acquisition du langage de très jeunes enfants sont capables de comprendre la signification émotionnelle de différents contours intonatifs. D'autre part, des études neuropsychologiques sur la perception de l'intonation indiquent que le traitement de l'intonation émotionnelle et le traitement de l'intonation linguistique peuvent être effectués de manière indépendante (v. Ross 1981, van Lanker & Sidtis 1992, Heilman, Bowers, Speedie & Coslett 1984), ce qui suggère que l'intonation pourrait contribuer à la communication émotionnelle indépendamment du contenu linguistique des énoncés.

Un contour de F0 est le plus souvent produit et interprété dans un contexte linguistique. En conséquence, il n'est sans doute pas opportun de chercher à identifier des contours spécifiques correspondant à des émotions spécifiques indépendamment du contexte linguistique. En revanche, certaines études – les travaux qui ont tenté d'isoler l'intonation et la qualité vocale, les recherches effectuées sur la compréhension prélinguistique des contours intonatifs émotionnels ou encore les études neuropsychologiques relatives au traitement de l'intonation – indiquent que certaines caractéristiques de l'intonation pourraient contribuer à communiquer l'émotion indépendamment des caractéristiques linguistiques des énoncés. Les caractéristiques qui joueraient ce rôle ne sont pas actuellement bien identifiées. Dans certaines études seul le contour de F0 est manipulé ce qui suggère implicitement que des aspects liés à l'ex-

cursion de la F0 sont conçus comme jouant un rôle particulièrement important dans le processus de communication émotionnelle. D'autres travaux considèrent conjointement un ensemble de modifications intonatives qui incluent des variations de F0, mais également de durée et d'intensité.

Dans l'approche présentée section 4, nous avons choisi de considérer uniquement l'excursion de la F0. Les contours de F0 correspondant à différentes expressions émotionnelles produites par des acteurs ont été stylisés de manière à pouvoir être systématiquement décrits et comparés. Sur la base de ces observations, des modifications systématiques de la F0 ont été appliquées à des expressions produites à l'aide du système de synthèse vocale par concaténation de diphones KALI (Morel 2001 : 193-221). La qualité émotionnelle de ces expressions de synthèse a été évaluée par des auditeurs en utilisant différentes procédures de jugement. Cette première approche permet d'évaluer dans quelle mesure le contour de la F0 est affecté par l'émotion exprimée et dans quelle mesure des modifications du contour de la F0 appliquées à une voix de synthèse parviennent à communiquer une impression émotionnelle sans recourir à des modifications du rythme (la durée totale et la durée relative des différents segments sont maintenues constantes) et de l'intensité.

A la section 5, une procédure plus proche de celle utilisée dans les études qui ont tenté d'isoler l'intonation et la qualité vocale est décrite. La F0, l'intensité et les durées de différents segments ont été extraites d'un ensemble d'expressions émotionnelles produites par des acteurs, puis ont été appliquées (copiées) sur des expressions produites par le système Kali. La qualité émotionnelle de ces expressions de synthèse qui reproduisent l'intonation naturelle des expressions émotionnelles avec la qualité vocale du système de synthèse a été évaluée par plusieurs groupes d'auditeurs. D'autre part, l'intonation des expressions originales produites par les acteurs a été "neutralisée" en appliquant le même rythme, intensité et contour de F0 à toutes les expressions par resynthèse. Ces expressions dont l'intonation est "neutralisées" mais qui gardent leur qualité vocale originale ont également été évaluées par plusieurs groupes d'auditeurs. Cette technique très prometteuse permet d'isoler véritablement sur le plan acoustique l'intonation (variations de F0, d'intensité et de durée) de la qualité vocale (modifications spectrales) et d'évaluer leurs contributions respectives à la communication émotionnelle.

2. LA PROSODIE EN SYNTHESE DE LA PAROLE

2.1. La prosodie : quoi, comment et pourquoi faire

Par définition, la prosodie est un phénomène suprasegmental, lié à la communication avec l'auditeur et caractérisé par la modification des paramètres de la parole (Lacheret 1999 : 11-12). Grâce à la prosodie, le locuteur améliore la transmission de l'information et communique en même temps son point de vue, ses intentions, voire ses sentiments. Plusieurs paramètres entrent en jeu : l'énergie, la tension des cordes vocales, l'articulation plus ou moins soignée, la vitesse de phonation, le degré de voisement, la présence éventuelle de souffle extraposé, ainsi que des phénomènes dus à des mouvements musculaires incontrôlés.

Au niveau du signal de parole, qui représente la vibration effectivement transmise – donc le support matériel de l'information, ces paramètres phonologiques se concrétisent de différentes manières :

- l'énergie se manifeste sous forme d'intensité, mais affecte également le voisement et le timbre (Liénard 1999 : 411-422),
- la tension des cordes vocales agit principalement sur la hauteur (F0), mais aussi sur le timbre (l'ensemble du conduit vocal se modifie),
- la qualité de l'articulation agit sur le timbre en modifiant l'amplitude et la forme de ses variations,
- la vitesse de phonation agit sur le paramètre temps, mais aussi sur l'articulation (plus soignée dans les parties lentes, plus relâchée dans les parties rapides) donc sur le timbre,
- le degré de voisement modifie le rapport source / souffle,
- le souffle extraposé se place entre les mots ou les groupes de mots et ajoute de l'information sur l'attitude ou l'émotion (irritation, mépris, peur, plaisir, etc.),
- les mouvements musculaires incontrôlés (dus généralement à l'émotion) se manifestent par des variations rapides et irrégulières de F0 et du timbre à l'intérieur même des phonèmes. Ils sont particulièrement difficiles à simuler en traitement du signal.

Pour les systèmes de synthèse de la parole, la prosodie est un enjeu important car elle représente une grande part de ce qui est humain dans la parole. La prosodie doit donner du relief à la voix et être jugée naturelle, sinon celle-ci paraît mécanique. Mais ce relief n'est pas disposé au hasard : pour une bonne intelligibilité, la prosodie

doit transmettre des informations syntaxiques (découpage, hiérarchisation), sémantiques et pragmatiques (degré de saillance et plus généralement structure communicative, Lacheret 2002 : 13-24). Pour une bonne expressivité, il est souhaitable également que la prosodie soit en mesure de communiquer des attitudes et des émotions.

Outre la difficulté de modéliser la prosodie, son implémentation concrète au niveau du traitement du signal se fait avec des outils encore trop rudimentaires : on sait modifier de façon assez satisfaisante l'intensité, la hauteur, la durée. Le degré de voisement et la génération de souffle nécessitent un modèle approprié. Mais le timbre, qui est affecté à des degrés divers par toutes les autres variations, est beaucoup plus difficile à modifier. Sa manipulation demande des connaissances difficilement accessibles et des outils puissants. Les expériences menées ici avec des outils capables de modifier la hauteur, l'intensité et la durée vont nous aider à préciser le rôle du timbre.

2.2. Place de la prosodie dans les systèmes de synthèse de la parole

Les systèmes de synthèse de la parole à partir du texte sont généralement composés de deux parties bien distinctes : (i) un module linguistique reçoit en entrée le texte et le convertit en une structure adaptée à la production de la parole ; (ii) un module phonétique est chargé de fabriquer le signal de parole à partir de cette structure (fig. 1).

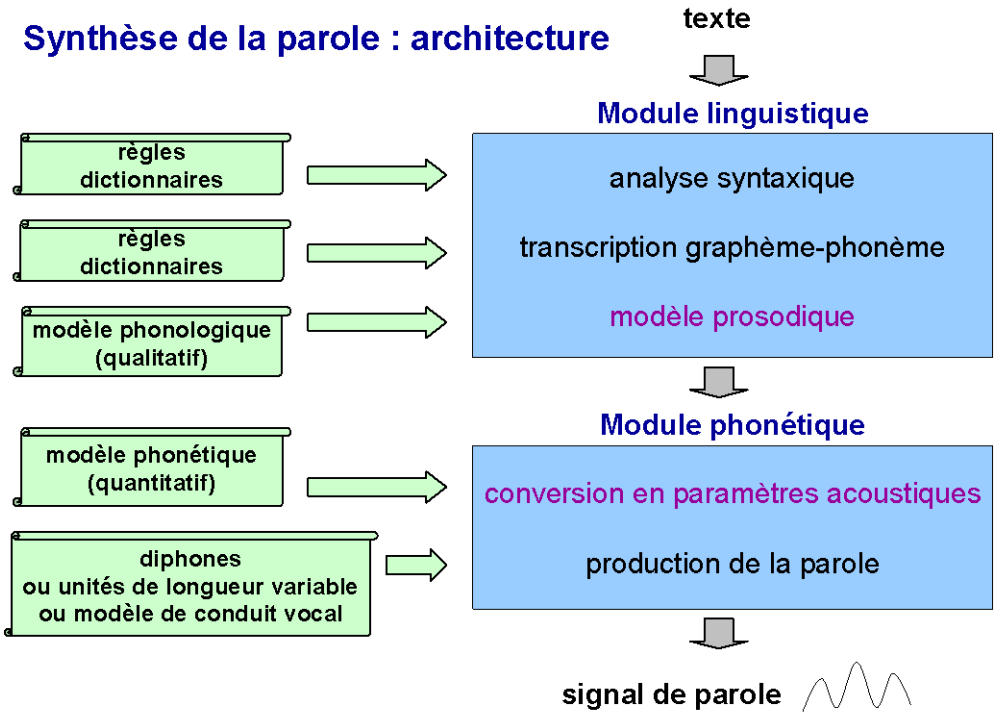


Figure 1 : architecture typique d'un système de synthèse de la parole

Le module linguistique procède généralement à une analyse syntaxique et effectue une transcription du texte en codes phonémiques. Un modèle prosodique qualitatif plus ou moins élaboré fournit une représentation de la structure communicative. A partir de cette structure, le module phonétique réalise une implémentation concrète de la prosodie sous forme de paramètres acoustiques, qui sont ensuite interprétés par le générateur de parole. Les paramètres manipulés sont généralement la hauteur, l'intensité et la durée.

Comme on le voit, la prosodie se situe à l'interface entre le module linguistique et le module phonétique. Nous nous intéresserons ici au module phonétique, et nous manipulerons les paramètres acoustiques de diverses manières pour tenter de simuler des attitudes et des émotions.

2.3. Codage de la prosodie dans Kali

Le système Kali utilise actuellement 1 seul triplet (hauteur, intensité, durée) par phonème, ce qui ne permet pas de simuler de variations à l'intérieur du même phonème. Dans un avenir proche, Kali utilisera 3 valeurs par voyelle pour la hauteur et l'intensité, atteignant un niveau de détail qui semble suffisant pour ces paramètres (Tournemire 1994 : 75-80).

En recopie prosodique (ou prosocopie), au contraire, le niveau de détail est trop élevé : nos outils recopient intégralement l'intonation d'un énoncé sur l'autre (fig. 2). De ce fait, la composante microprosodique est recopiée alors qu'elle appartient au niveau segmental et est propre à chaque locuteur. Là encore, nous pensons que le passage à 3 valeurs par voyelle devrait correspondre au bon degré de détail.

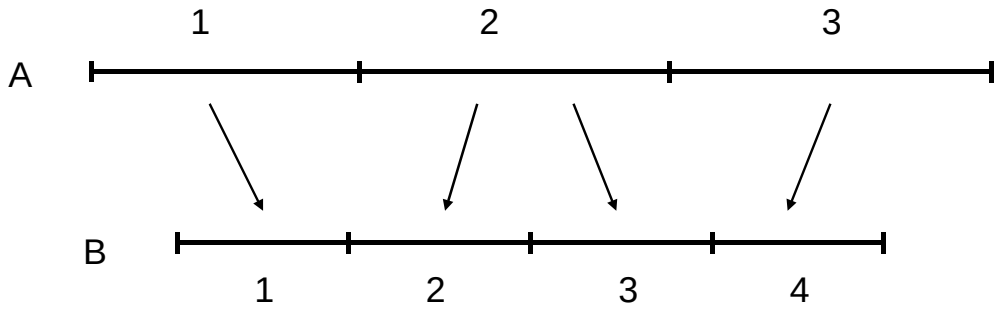


Figure 2 : recopie intégrale de l'intonation : modification d'un diphone

Dans cet exemple, le diphone A, formé de 3 périodes, est modifié conformément au diphone B, formé de 4 périodes et doit rentrer dans son gabarit. La taille des périodes est modifiée (voix plus aiguë) et A2 est dédoublée en B2 et B3. La durée devient celle de B. Une fois l'amplitude modifiée, l'intonation de B se trouve recopiée intégralement sur le signal A.

2.4. Variation de la hauteur et de la durée dans Kali

Comme dans la plupart des systèmes actuels, la méthode utilisée pour modifier la fréquence fondamentale et la durée en restant dans le domaine temporel consiste à additionner des fenêtres alignées sur la période. Cette méthode, déjà expérimentée depuis plusieurs dizaines d'années de façon parfois rudimentaire (Morel 1981 : 24-27) s'est répandue dans les années 90 sous le nom de TD-PSOLA, algorithme créé par le CNET (Hamon 1989 : 238-241). Elle présente le gros avantage de nécessiter peu de calculs, ce qui permet de l'appliquer en temps réel pendant la production de la parole pour concrétiser la prise en compte des variations prosodiques, tout en conservant au mieux la qualité du timbre original.

En pratique, le signal de période courante T est découpé en signaux élémentaires de largeur 2T, résultant du produit du signal d'origine par une fenêtre en cloche, de largeur 2T, centrée sur la partie de plus forte énergie de chaque période. Si les fenêtres sont sinusoïdales (fenêtres de Hanning), la somme des signaux élémentaires, qui se recouvrent mutuellement sur la largeur d'une période, reproduit le signal d'origine (fig. 3). La forme des fenêtres diffère selon les auteurs. Après des essais perceptifs sur plusieurs types de fenêtres, nous avons constaté que c'était moins la forme de la fenêtre que sa position par rapport au signal qui était importante, et nous avons adopté la forme sinusoïdale, qui présente l'avantage d'une bonne répartition de l'énergie, permettant d'étendre le traitement des parties voisées du signal à celui des parties sourdes.

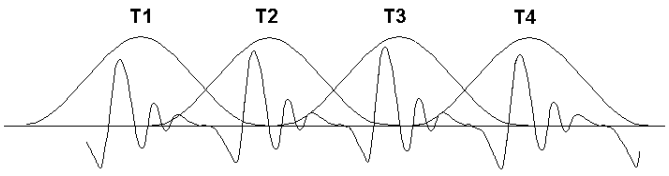


Figure 3 : juxtaposition des signaux élémentaires après fenêtrage

La diminution de la période consiste à resserrer les signaux élémentaires tout en diminuant conjointement la largeur des fenêtres sinusoïdales, de façon à conserver une répartition uniforme de l'énergie. L'augmentation de la période correspond à l'opération inverse, à ceci près qu'il n'est pas souhaitable de dilater les fenêtres, car, trop larges, elles contiendraient alors des traces non négligeables de la périodicité d'origine, créant des phénomènes de réverbération préjudiciables à la qualité de la parole. Les fenêtres sont donc écartées sans modification, ce qui ne détériore pas significativement la qualité de la voix synthétique. En revanche, les parties non ou peu voisées (en pratique inférieures à un certain taux de voisement) dont l'énergie est répartie plus uniformément, doivent bénéficier de la dilatation des fenêtres, évitant ainsi

un phénomène de périodisation, qui serait lui aussi préjudiciable à la qualité de la voix.

La modification de durée (ou de son inverse, la vitesse de phonation) est effectuée en dupliquant ou en retirant des signaux élémentaires. Le problème de largeur des fenêtres se pose moins ici, les fenêtres à superposer étant de tailles voisines. En conséquence, la modification de durée par les méthodes temporelles affecte très peu la qualité du signal résultant.

3. APPLICATION PRELIMINAIRE : CORPUS RADIOPHONIQUE

Une expérience d'échange d'intonation est réalisée à partir d'un corpus radiophonique – interview de Benoîte Groult par Roselyne Fayard – transcrit sous forme de texte, pour évaluer séparément l'influence de l'intonation et de la qualité vocale sur la qualité générale du dialogue. Il s'agit (i) de recopier l'intonation du dialogue réel sur la voix synthétique et inversement (ii) de recopier l'intonation du modèle de Kali sur la voix réelle des locuteurs. Seule une partie du corpus (la plus animée) a été traitée.

L'expérience (i) a pour but de mettre en évidence la qualité de la voix synthétique et son aptitude à suivre une intonation donnée et à exprimer les attitudes qui y sont codées. Dans l'ensemble, les variations imposées à la voix synthétique sont bien acceptées par le système. Des défauts de timbre sont décelables car, comme nous l'avons vu, les variations des paramètres intonatifs devraient être corrélées à des variations de timbre. Mais ce qui frappe le plus, c'est le manque de corrélation entre la vitesse de phonation et la qualité articulatoire, phénomène très sensible sur la première phrase traitée : elle est rapide et paraît articulée à l'excès, donnant une impression d'impatience. Un bon modèle prosodique devrait donc pouvoir jouer sur la qualité articulatoire.

En ce qui concerne la recopie des attitudes, le résultat est très flatteur : on retrouve bien dans le dialogue l'interrogation, l'empathie, l'évidence, l'indignation, la lassitude, mais il est vrai que le contenu linguistique nous y aide… Seules des expériences menées avec rigueur pourraient confirmer cette impression. Néanmoins c'est un premier résultat à rapprocher du second (sur les émotions) très contrasté par rapport à celui-ci. La qualité obtenue montre aussi le progrès qui reste à accomplir pour obtenir une intonation naturelle et dynamique dans Kali.

L'expérience (ii) a pour but de mettre en évidence la qualité du modèle intonatif de Kali. Le résultat, dans l'ensemble moins bon que le précédent, montre un modèle trop stéréotypé, manquant de relief, dans lequel l'absence de modèle énonciatif se fait sentir. L'intonation doit donc être améliorée en priorité. La voix devra cependant progresser elle aussi pour mieux s'adapter aux variations prosodiques. Mais ce qui frappe, là encore, c'est le problème de la vitesse de phonation : dans cette expérience, le phé-

nomène inverse du précédent se produit, à savoir une impression d'articulation négligée dans certaines parties, surtout dans la première phrase traitée. Celle-ci est prononcée rapidement par le locuteur et le modèle la ralentit, l'auditeur s'attend alors à une meilleure qualité articulatoire. Ce phénomène vient biaiser l'expérience car si plusieurs stratégies prosodiques sont possibles, l'originale se trouve favorisée ; le modèle prosodique est donc jugé trop sévèrement. Quant à la qualité des voix naturelles dans les parties ou la prosodie est acceptable, elle nous donne une idée du progrès qui reste à accomplir sur la voix synthétique pour la rendre plus agréable et naturelle.

4. APPLICATION A LA SYNTHESE DES EMOTIONS : STYLISATION DE LA FREQUENCE FONDAMENTALE

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 émotionnelles, 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.

4.1. Analyse d'un corpus de 144 enregistrements

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 et 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 gött 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 ('exalt'). La fréquence fondamentale a été extraite par auto-corrélation pour chacune des 144 expressions émotionnelles à l'aide du logiciel PRAAT (Boersma & Weenink 1996).

Dix points de chaque contour de F0 devaient en principe être 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. 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". 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". La figure 4 présente une illustration de ce codage pour une expression émotionnelle. Le contour de F0 original est représenté en gris, le contour stylisé surimposé en vert/noir. Le point 8 ('3min') est absent dans cette expression.

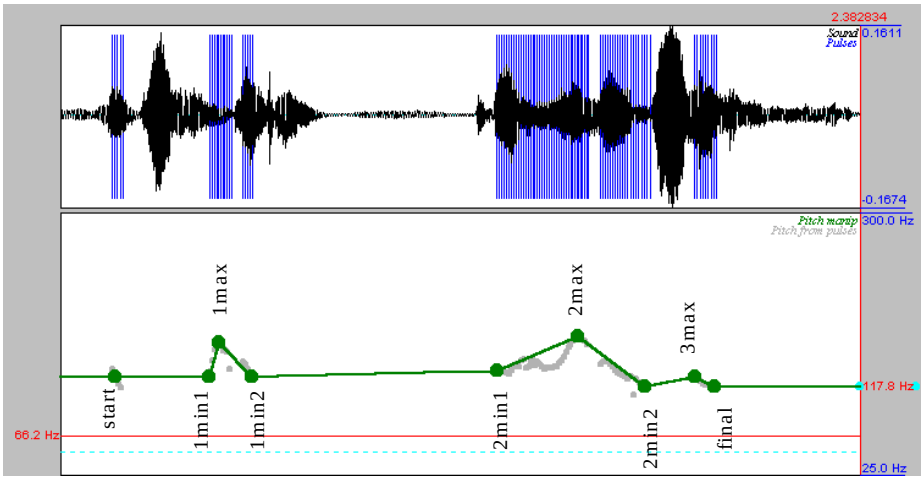


Figure 4 : illustration du codage du contour de F0 pour une expression émotionnelle

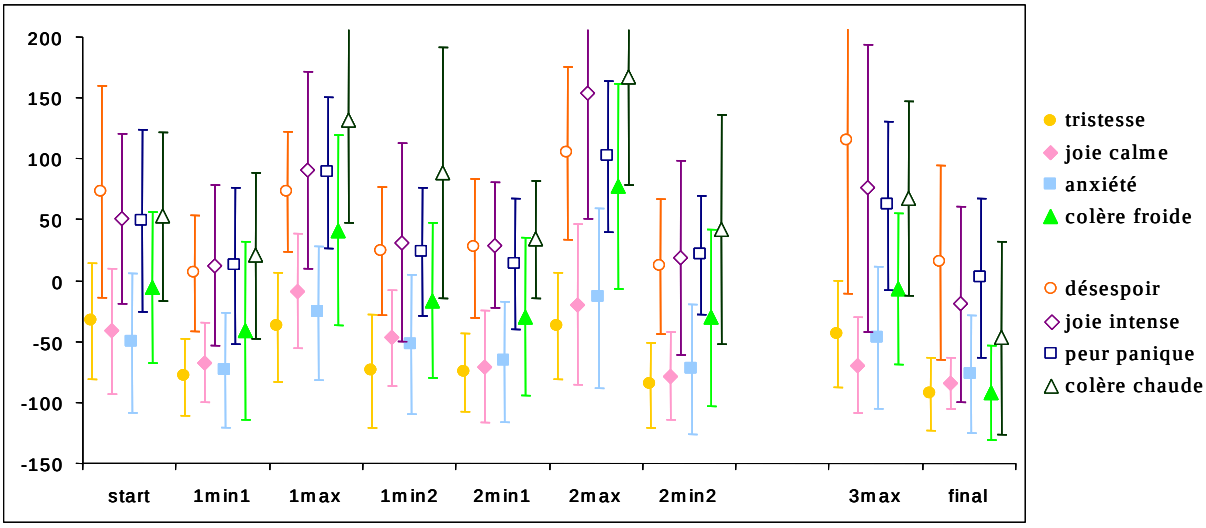
Pour une partie assez importante des expressions, cette configuration n'est pas réalisée. Pour le premier point observé (première syllabe de chaque "phrase"), le point peut être présent ou absent. Pour les trois autres parties sur lesquelles des excursions sont attendues ('acc1', 'acc2' et 'final'), 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. Le tableau 1 représente le nombre d'excursions de chaque type observé en fonction de l'émotion exprimée.

Tableau 1: Types d'excursion de la F0, nombre par segment et par émotion exprimée.

segment	excursion F0	anx	joie	rage	irrit	paniq	trist	exalt	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	4	4	5	6	1	9	6	4.88
	montée	0	2	1	0	1	3	1	0	1.00
	descente	8	9	11	11	9	6	6	10	8.75
	absent	6	3	2	2	2	8	2	2	3.38

Les résultats présentés dans le tableau 1 indiquent que les émotions exprimées n'affectent pas le type d'excursion codé pour les quatre segments considérés ; aucune différence importante n'apparaît pour aucune des émotions exprimées.

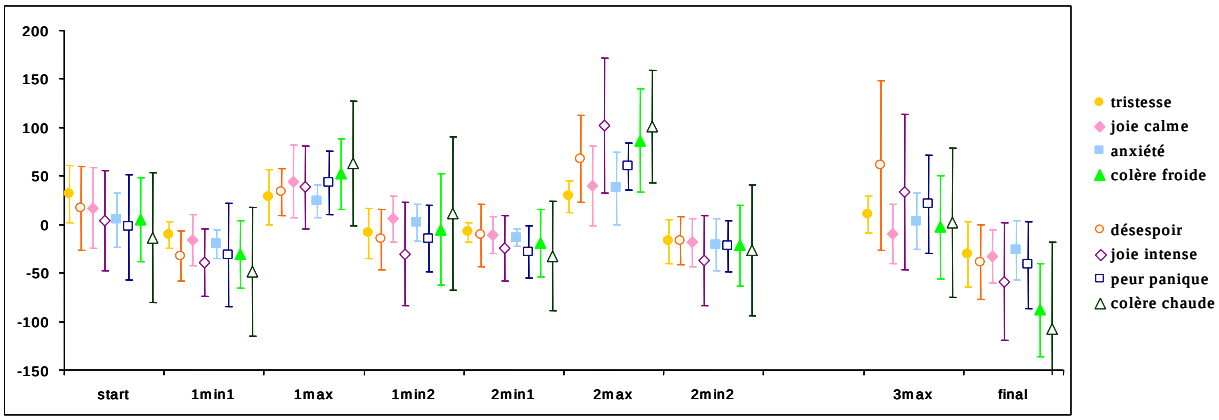
Etant donné la quantité inégale de points de F0 relevés pour les différents locuteurs, il s'est avéré nécessaire de contrôler le niveau de F0 propre à chaque locuteur avant d'évaluer si l'amplitude des excursions de F0 diffère en fonction des émotions exprimées. A cette fin, la F0 moyenne de chaque locuteur a été calculée sur la base de 112 expressions produites par chaque locuteur. 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 47 des 144 expressions et seulement cinq fois pour la première "phrase".



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, une 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. D'autre part, les différences qui apparaissent dans ce graphique semblent liées surtout au niveau d'activation sous-jacent aux émotions exprimées : les moyennes sont plus faibles pour les émotions faiblement activées (représentées par les symboles pleins) et plus élevées pour les émotions fortement activées (représentées par les symboles vides).

Plus généralement, il semble que la mesure des différents points n'ajoute que peu d'information à une mesure plus globale de la F0 qui permet également de différencier les émotions faiblement activées des émotions fortement activées. Afin de tester cette affirmation, la F0 moyenne de chaque expression a été régressée sur l'ensemble des points de F0 relevés pour toutes les expressions. Le graphique 2 représentent les moyennes et les écarts-types des résidus de cette régression pour chaque point de F0 relevé en fonction du type d'émotion exprimée.



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

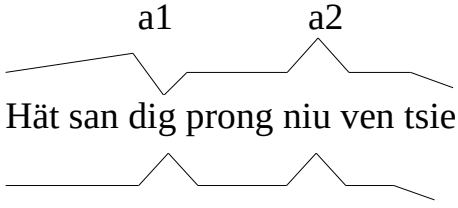
Ce graphique permet d'observer qu'après le contrôle de la F0 moyenne, une différence en fonction de l'émotion exprimée pourrait encore apparaître pour le point '2max' et le point 'final'. Le deuxième maximum présente des valeurs moyennes plus élevées pour la colère chaude, la colère froide et la joie intense que pour les autres émotions exprimées. En ce qui concerne le point 'final', la moyenne pour la colère chaude et, dans une moindre mesure, la moyenne pour la colère froide sont plus faibles que les moyennes pour les autres émotions exprimées. Cette observation permet de formuler l'hypothèse que les expressions correspondant à certaines émotions – en l'occurrence les deux formes de colère et la joie intense – pourraient présenter un deuxième accent plus proéminent que le premier. D'autre part, la chute finale du contour permettrait également de différencier en partie les émotions exprimées, alors que le niveau général de la F0 reflèterait essentiellement l'activation sous-jacente à l'émotion exprimée.

L'existence de différences au niveau de l'amplitude des excursions de F0 pour différents types d'émotions est également confirmée par l'examen des pentes relevées (les pentes de chaque "montée" et de chaque "descente" de F0 ont été calculées en divisant la différence entre les minima et les maxima de F0 contigus par la durée de ces excursions). 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.

Il semble donc que des caractéristiques du contour de F0 – telles que l'amplitude des excursions, la proéminence relative de différentes excursions ou encore la chute finale du contour – sont affectées par le type d'émotion exprimé. 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 (Kali). La qualité émotionnelle perçue de ces expressions a été ensuite évaluée dans des tests de perceptions. La procédure utilisée est décrite ci-dessous.

4.2. Synthèse de 96 expressions avec stylisation de F0

Conformément à l'analyse du corpus ci-dessus, la courbe F0 des deux énoncés prononcés en parole de synthèse est manipulée systématiquement selon une stylisation simple :



(1) Hät san dig prong niu ven tsie

(2) Fi göt laich jean kill gos terr

Deux niveaux de base ("bas" et "haut") et deux amplitudes d'accents ("petit" et "grand") appliquées à deux accents (a1 et a2), ainsi que trois mouvements sur la syllabe finale ("montant", "descendant" et "plat") sont appliqués à ces énoncés, avec deux voix différentes. Au total, 96 expressions sont créées : 2 énoncés x 2 voix x 2 niveaux x 2 accents1 x 2 accents2 x 3 finales.

Les valeurs ont été choisies aussi fortes que possible pour une action optimale, tout en restant dans un registre acceptable. Comme des variations fortes (supérieures à une octave) altèrent le timbre, nous avons dû rester en deçà des valeurs extrêmes relevées sur les productions des acteurs. Les valeurs des extremums en tons sont les suivantes :

niveau de base	accent1	accent2	finale
0	3	3	-4
4	6	6	0
			4

La première impression qui se dégage à l'écoute des 96 énoncés est que ceux-ci semblent exprimer certaines attitudes (affirmation, insistance, interrogation) mais pas les émotions prévues (tristesse, joie, peur, colère). L'évaluation perceptive (section 6) confirme cette impression.

5. APPLICATION A LA SYNTHESE DES EMOTIONS : PROSOCOPIE CROISEE SUR 16 ENONCES EMOTIONNELS

Etant donné le faible résultat obtenu avec des variations simples de hauteur, nous voulons ici savoir si un modèle intonatif "parfait" (mais sans timbre) permet de simuler les huit émotions de référence. Nous jouons donc sur les trois paramètres hauteur, intensité et durée, dont nous recopions intégralement les variations sur la voix synthétique. La manipulation inverse (intonation plate recopiée sur les énoncés originaux) permet en principe de retrouver la partie des émotions qui est portée par le timbre.

5.1. Synthèse des 32 expressions

Pour cette tentative, 16 expressions émotionnelles (8 types d'émotions, 2 énoncés) ont été sélectionnées parmi les 144 expressions utilisées pour la stylisation des contours décrite à la section 4.1. Des expressions dont la qualité émotionnelle a été particulièrement bien reconnue lors dans un test perceptif préalable (Bänziger & Scherer 2001) ont été choisies. Les deux énoncés (1. "hät san dig prong nju ven tsi", 2. "fi götт laich jean kill gos terr") sont donc prononcés par différents acteurs et communiquent 8 ty-

pes 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').

La première remarque qui s'impose est que ces énoncés posent des problèmes techniques : les signaux sont beaucoup plus difficiles à analyser que des signaux de parole lue ou même de parole spontanée. La principale difficulté consiste à déterminer la fréquence fondamentale, même manuellement. Les signaux chargés de beaucoup d'émotion sont perturbés (contractions, souffle) et peu périodiques, rendant le marquage parfois aléatoire. Les résultats sont donc à interpréter avec prudence.

Malgré ces réserves, dès la première écoute, la voix synthétique avec intonation émotionnelle – comparée à celle des énoncés originaux – paraît dénuée d'émotion, bien que d'aspect naturel. Elle semble, comme précédemment, exprimer des attitudes modérées. La comparaison des résultats entre eux donne un meilleur résultat : par contraste, on reconnaît quelques énoncés comme étant un peu plus gais ou un peu plus tristes. La manipulation inverse fournit des énoncés nettement plus émotionnels : la majeure partie de ce type d'émotions serait donc codée dans le timbre. L'évaluation perceptive (section suivante) confirme et précise cette impression.

6. EVALUATION PERCEPTIVE DES EXPRESSIONS OBTENUES

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. Les expressions naturelles (16 expressions), les expressions modifiées par synthèse (32 expressions produites par la prosocopie croisée, v. section 5.1) et les expressions de synthèse dont seul le contour de F0 a été systématiquement modifié (96 expressions, v. section 4.2) ont été présentées durant la même session dans un ordre aléatoire différent pour chaque auditeur. Les auditeurs ont évalué pour chaque expression présentée:

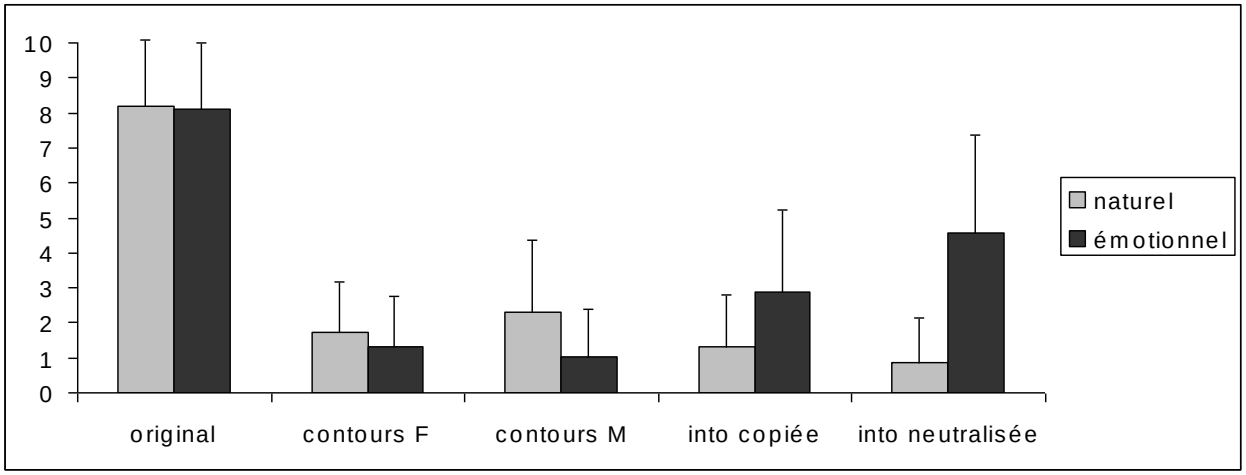
- A quel degré l'expression produit une impression naturelle (versus artificielle).
- A quel degré l'expression parvient à communiquer une émotion ou une attitude.
- A quel point cette émotion/attitude est positive (versus négative).
- Quel est le degré d'intensité de cette émotion/attitude

Les jugements sont effectués sur des échelles visuelles analogues, les réponses sont enregistrées sur des échelles continues de 0 à 10.

Les corrélations intraclasses calculées (indices de fidélité des réponses entre les auditeurs) indiquent que les auditeurs ne parviennent pas à évaluer de manière fidèle les caractéristiques ci-dessus 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 prosocopie croisée, 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.

Dans ce mode d'évaluation, où toutes les expressions sont présentées dans une même session, les jugements sont en fait surtout fonction du type d'expression présenté (v. graphique 3). En comparaison directe avec les expressions de synthèse, les expressions naturelles sont jugées très naturelles et très émotionnelles ; alors que les expressions de synthèse (prosocopie et manipulation systématique des contours) sont jugées beaucoup moins naturelles et beaucoup moins émotionnelles. L'absence de fidélité pour les jugements relatifs aux expressions dont les contours de F0 ont été systématiquement modifiés, pourrait être due également à cet effet de contraste : 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. Des différences qui pourraient éventuellement apparaître entre les expressions dont le contour de F0 a été modifié seraient en conséquence gommées dans ce contexte.



Graphique 3: moyennes des jugements relatifs au degré naturel et émotionnel des expressions

Les barres dans ce graphique représentent les jugements moyens pour les expressions naturelles (original), pour les contours de F0 appliqués à la voix féminine (contours F), pour les contours appliqués à la voix masculine (contours M), pour les expressions qui conservent l'intonation naturelle appliquée sur la voix de synthèse (into copiée), pour les expressions qui conservent la qualité vocale naturelle avec l'intonation du modèle (into neutralisée). On observe que les expressions produites par la prosocopie croisée (en particulier les expressions qui conservent la qualité vocale naturelle – into neutralisée) sont perçues comme plus émotionnelles, mais pas comme plus naturelles, que les contours de F0 stylisés qui ont été appliqués à la voix de synthèse.

Afin de pallier au problème du manque de fidélité des réponses, une deuxième procédure de jugement a été mise en place: 4 dimensions émotionnelles représentant l'intensité de colère, de joie, de tristesse et de peur perçues (la figure 5 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. Quinze auditeurs 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 est aléatoire). 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 (into copiée), ainsi que les 16 expressions qui ont été produites en imposant l'intonation du système de synthèse sur les 16 expressions naturelles (into neutralisée). L'ordre de présentation de ces deux ensembles d'expressions a été défini aléatoirement pour chaque auditeur.

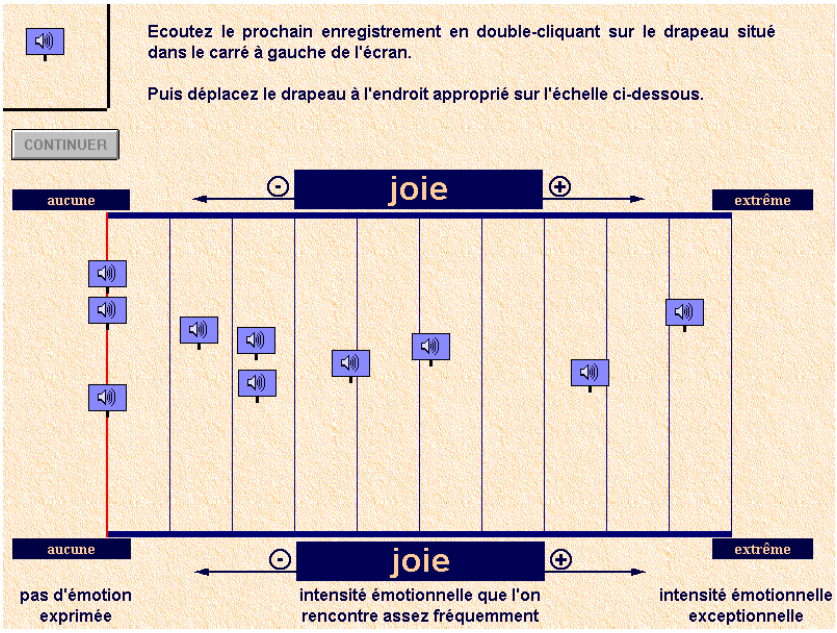


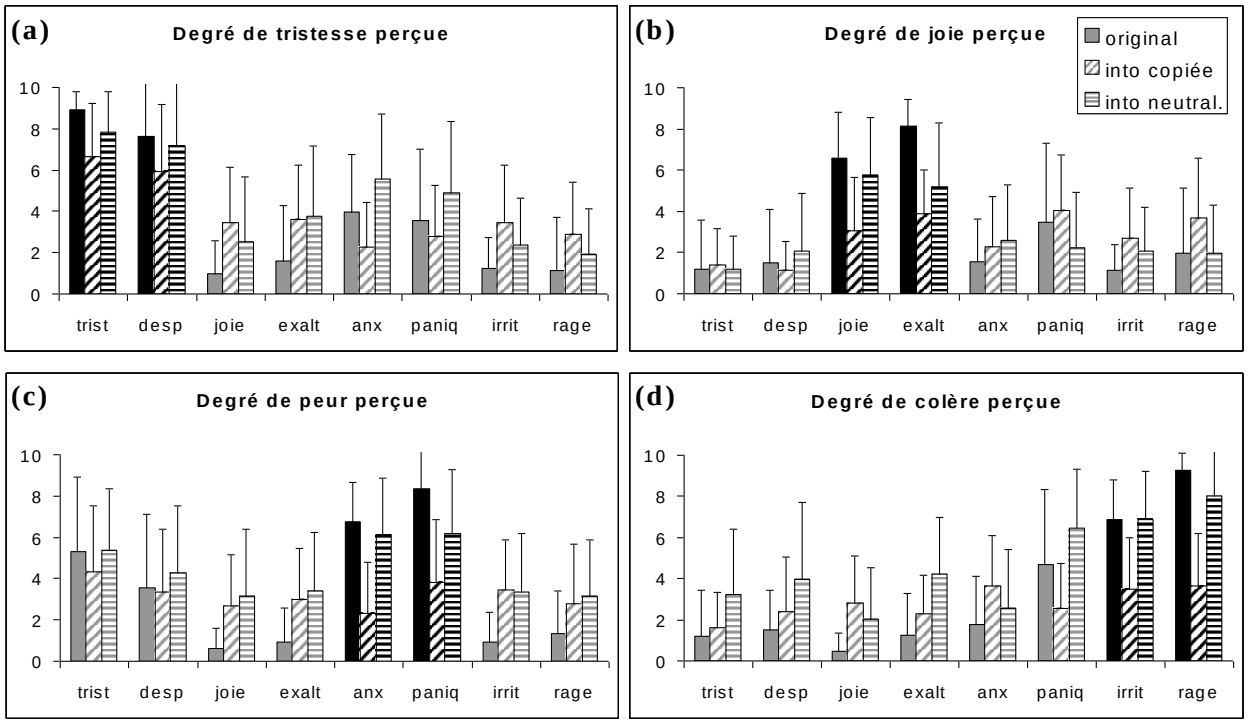
Figure 5: Exemple d'une échelle utilisée pour évaluer l'intensité de l'émotion perçue dans les expressions (les expressions sont représentées graphiquement par les "drapeaux" bleus)

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 2 (r = single measure intraclass correlation ; R = average measure intraclass correlation ; N = 15) 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 dont l'intonation originale et préservée et pour les 16 expressions dont la qualité vocale est préservée.

Tableau 2: Corrélations intraclasses pour les jugements d'intensité de joie, peur, colère et tristesse.

Expressions	joie		peur		colère		tristesse	
	r	R	r	R	r	R	r	R
contours	.35	.89	.61	.96	.04	.36	.05	.42
into copiée	.25	.83	.24	.82	.32	.88	.33	.88
into neutralisée	.41	.91	.36	.90	.40	.91	.39	.91

Pour les expressions produites par la prosocopie croisée ('into copiée' et 'into neutralisée'), les jugements des auditeurs sont relativement bien corrélés pour les 4 dimensions émotionnelles évaluées. Les graphiques 4a-4d représentent les moyennes et les écarts-types par émotion exprimée pour les expressions naturelles (barres sombres) qui ont été évaluées séparément par un autre groupe de 16 auditeurs, pour les expressions dont l'intonation est préservée (barres rayées diagonalement) et pour les expressions dont la qualité vocale est préservée (barres rayées horizontalement).

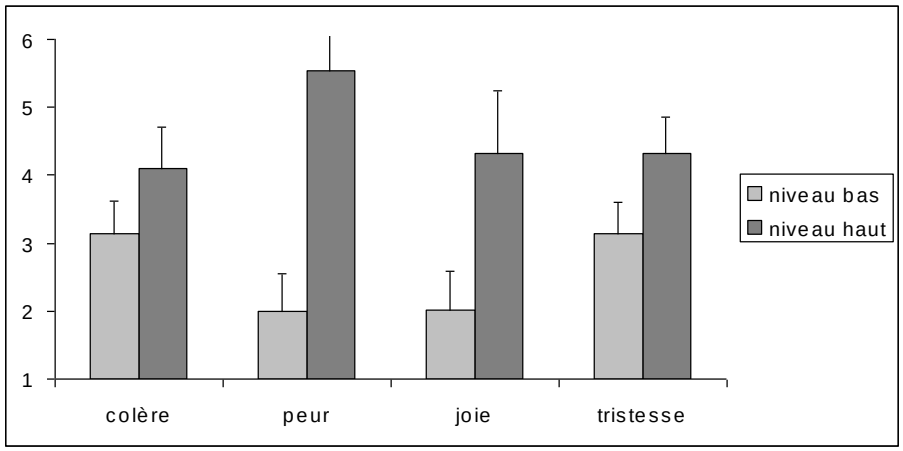


Graphiques 4a-4d: moyennes et écarts-types de l'intensité émotionnelle perçue: pour la tristesse (a), la joie (b), la peur (c), la colère (d) par émotion exprimée, pour les expressions naturelles, les expressions dont l'intonation est préservée et les expressions dont l'intonation est neutralisée

Les graphiques 4a à 4d indiquent que la perception de l'émotion exprimée est dégradée dans une plus forte mesure pour les expressions dont l'intonation est préservée (barres rayées diagonalement) que pour les expressions dont la qualité vocale est préservée (barres rayées horizontalement). En fait seule la tristesse reste encore assez bien communiquée dans les expressions qui ont été produites en copiant l'intonation des expressions naturelles sur la voix de synthèse. Pour les expressions produites en

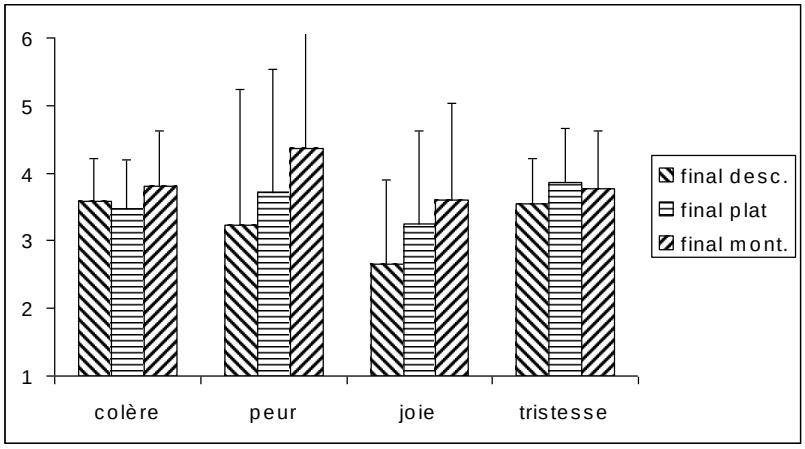
"neutralisant" l'intonation et en conservant la qualité vocale, les 4 types d'émotions exprimées restent relativement mieux reconnues.

En ce qui concerne les expressions dont le contour de F0 a été systématiquement modifié, la manipulation qui affecte le plus les jugements émotionnels est le niveau de la F0 (v. graphique 5). L'intensité émotionnelle pour les 4 catégories évaluées (colère, peur, joie, tristesse) a été jugée plus importante pour les expressions dont le niveau des contours a été élevé de 4 tons. L'intensité de peur évaluée est plus particulièrement affectée par le niveau de la F0, l'influence du niveau de la F0 sur les jugement de colère et de tristesse est beaucoup plus limitée.



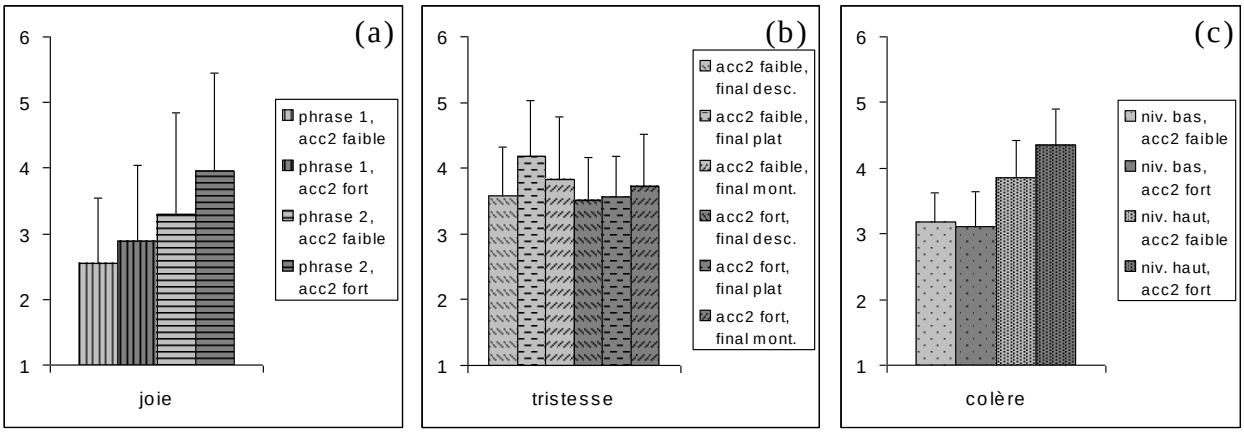
Graphique 5: Intensité de colère, de peur, de joie et de tristesse perçue en fonction du niveau de la F0 imposé sur les expressions de synthèse

Dans une moindre mesure, les jugements émotionnels sont également affectés par le mouvement final de la F0. 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 (v. graphique 6). Les jugements moyens concernant l'intensité de colère et de tristesse perçues ne sont pas influencés par le mouvement final des contours de F0.



Graphique 6: Intensité de colère, de peur, de joie et de tristesse perçue en fonction du mouvement final de la F0 (descendant, plat, montant) imposé sur les expressions de synthèse

Les effets d'interaction entre les différents aspects des expressions qui ont été modifiés (niveau, 1er "accent", 2ème "accent", mouvement final, phrase) sont nombreux et constituent les principaux effets distinctifs pour les 4 catégories émotionnelles évaluées. Les graphiques 7a-7c présentent trois exemples d'effets d'interaction entre la proéminence du 2ème "accent" et la phrase (7a), le mouvement final (7b), le niveau (7c). La phrase 1 avec un deuxième "accent" peu proéminent et perçue comme exprimant une joie moins intense que la phrase 2 avec un deuxième "accent" plus proéminent. Les expressions avec un mouvement final plat et un 2ème "accent" faible sont perçues comme plus tristes que les expressions avec un mouvement final descendant et un 2ème "accent" proéminent. Les expressions avec un niveau de F0 haut et un 2ème "accent" fort sont perçues comme exprimant une intensité de colère plus forte que les expressions avec un niveau de F0 bas.



Graphique 7a-7c: Effets d'interaction entre la proéminence du 2ème accent et: (a) la phrase sur l'intensité de joie perçue, (b) le mouvement final sur l'intensité de tristesse perçue, (c) le niveau de F0 sur l'intensité de colère perçue

7. DISCUSSION ET CONCLUSION

Les expériences décrites dans ce chapitre montrent que les paramètres hauteur, intensité et durée ne parviennent pas toujours à exprimer de façon satisfaisante les attitudes et émotions. Il semble que certaines émotions (instinctives, involontaires, incontrôlées) nécessitent le recours à des variations spécifiques de timbre, l'intonation ne jouant qu'un rôle secondaire d'accompagnement. On peut objecter que la production des acteurs – sur laquelle nous nous appuyons – est volontaire et contrôlée. Mais précisément, il est fait appel à des acteurs pour suppléer au fait que la mise en situation émotionnelle de sujets poserait des problèmes éthiques : les acteurs, eux, sont capables de s'imprégner des émotions qu'ils jouent et de simuler leur caractère incontrôlé. L'artéfact reste donc limité, d'autant plus que les meilleures productions ont été sélectionnées par des expériences perceptives.

Les données ne permettent pas d'affirmer que l'intonation n'est pas utilisée par les auditeurs pour reconnaître les émotions : d'une part, les expressions dont l'intonation a été neutralisée sont jugées moins émotionnelles que les expressions naturelles et, d'autre part, la discrimination des émotions exprimées est dégradée pour ces expressions. Mais, il est apparu que l'intonation seule ne semble pas parvenir à communiquer une impression émotionnelle. Les contours stylisés et l'intonation copiée sont perçus comme non-émotionnels et l'émotion exprimée n'est presque plus reconnue pour les expressions dont l'intonation a été copiée sur la voix de synthèse. A l'inverse, le timbre isolé réussit encore à transmettre une partie de l'information émotionnelle.

Ce travail présente d'autres limites d'ordre technique : d'une part le faible nombre d'expressions utilisées (16 expressions soit 2 par émotion exprimée), insuffisant pour généraliser, d'autre part la difficulté à neutraliser l'intonation, difficulté dont la conséquence est probablement un reste d'intonation émotionnelle dans les énoncés neutralisés.

Nous avons montré, en tout cas, à quel point il serait illusoire de prétendre simuler des émotions sans tenir compte du timbre. La modélisation de ce dernier avec les méthodes habituelles semblant difficile, deux voies nous paraissent possibles : (i) la modélisation du conduit vocal, si elle parvient un jour à produire des résultats de qualité en temps réel, pourra prendre en compte les phénomènes physiques liés à l'émotion. (ii) Les nouvelles méthodes de synthèse vocale par sélection d'unités dans des grands corpus (Bozkurt 2002 : 1-4) présentent déjà l'avantage de produire une parole mieux adaptée aux variations intonatives (par exemple, les parties rapides – donc peu articulées – sont choisies préférentiellement dans des parties rapides – donc peu articulées – du corpus car leur score est meilleur). Il n'est pas inconcevable de penser que l'utilisation de corpus émotionnels – moyennant l'utilisation de ressources de parole compressée considérables – permettra à terme de simuler les émotions de façon satisfaisante.

Michel MOREL
Laboratoire CRISCO - CNRS
Université de Caen
14032 CAEN CEDEX
morel@crisco.unicaen.fr

Tanja BÄNZIGER
FAPSE/Université de Genève
40 bv du Pont-d'Arve
1205 Genève
Tanja.Banziger@pse.unige.ch

REFERENCES BIBLIOGRAPHIQUES

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

BÄNZIGER T. & K.R. SCHERER. 2001. « Relations entre caractéristiques vocales perçues et émotions attribuées », Actes des Journées Prosodie, Grenoble, 10-11 octobre 2001.

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

BOZKURT B., T. DUTOIT, R. PRUDON, C. d'ALESSANDRO & V. PAGEL. 2002. « Improving quality of MROLA synthesys for non-uniform units synthesis », Proceedings of the IEEE TTS 2002 Workshop, Santa Monica, Septembrer 2002, 1-4.

ELFENBEIN H.A. & N. AMBADI. 2002. « On the universality and cultural specificity of emotion recognition: A meta-analysis », *Psychological Bulletin* 128, 203-235.

FONAGY I. & K. MAGDICS. 1963. « Emotional patterns in intonation and music », *Zeitschrift für Phonetik* 16, 293-326.

HAMON C., E. MOULINES & F. CHARPENTIER. 1989. « A diphone synthesis system based on time-domain prosodic modifications of speech », International Conference on Acoustics, Speech and Signal Processing (ICASSP), T. Durani éd., Glasgow, 238-241.

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

JOHNSTONE T. & K.R. SCHERER. 2000. « Vocal communication of emotion », dans M. LEWIS & J. HAVILAND-JONES (dir.), *Handbook of Emotions, Second Edition*, New York: Guilford Press, 220-235.

JUSLIN P. N. & P. LAUKKA. 2001. « Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion », *Emotion* 1, 381-412.

KLASMEYER G. 1999. « Akustische Korrelate des stimmlich emotionalen Ausdrucks in der Lautsprache », *Forum Phoneticum 67*, Frankfurt am Main: Hector.

LACHERET-DUJOUR A. & F. BEAUGENDRE. 1999. *La prosodie du français*, Paris: éditions du CNRS, 11-12.

LACHERET-DUJOUR A. 2002. « Modélisation prosodique du français parlé : analyse de la substance, représentation formelle, interprétation symbolique », Dossier d'habilitation à diriger des recherches, Université de Paris X, Nanterre, juillet 2002.

LADD D.R., K.E.A. SILVERMAN, F. TOLKMITT, G. BERGMANN & K.R. SCHERER. 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.

LIEBERMAN P. & S.B. MICHAELS. 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.

LIENARD J.S. & M.G.D. BENEDETTO. 1999. « Effect of vocal effort on spectral properties of vowels », *Journal of the Acoustical Society of America* 106, 411-422.

MOREL M. 1981. « Synthèse vocale par raccordement de segments d'oscillogrammes », *Revue d'Acoustique du GALF* 56, 24-27.

MOREL M. & A. LACHERET-DUJOUR. 2001. « Le logiciel de synthèse vocale Kali : de la conception à la mise en œuvre », dans Ch. D'ALESSANDRO (dir.), *Traitement Automatique des Langues* 42, Paris: Hermès, 193-221.

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

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

PAPOUSEK M. 1994. *Vom Schrei zum ersten Wort: Anfänge der Sprachentwicklung*, Bern: Verlag Hans Huber.

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. 2003. « Vocal communication of emotion: A review of research paradigms », *Speech Communication* 40, xx-xx.

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

SCHERER K.R., D.R. LADD & K.E.A. SILVERMAN. 1984. « Vocal cues to speaker affect: Testing two models », *Journal of the Acoustical Society of America* 76, 1346-1356.

TOURNEMIRE S. 1994. « Recherche d'une stylisation extrême des contours de F0 en vue de leur apprentissage automatique ». Journées d'Etudes sur la Parole, Trégastel, France, 75-80.

ULDALL E. 1964. « Dimensions of meaning in intonation », dans D. ABERCROMBIE, D.B. FRY, P.A.D. MACCARTHY, N.C. SCOTT & J.L.M. TRIM (dir.), *In honour of Daniel Jones: papers contributed on the Occasion of his Eightieth birthday, 12 september 1961*, London: Longman, 271-279.

VAN LANKER D. & J.J. SIDTIS. 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.