L'état émotionnel du locuteur: facteur négligé mais non négligeable pour la technologie de la parole

K. R. Scherer, T. Johnstone, J. Sangsue

FPSE, Université de Genève
9, Route de Drize, 1227 Carouge, Suisse
e-mail: scherer@uni2a.unige.ch

"Il y a autant de mouvements dans la voix qu'il y a de mouvements dans l'esprit, et l'esprit est profondément affecté par la voix."

Cicéron, De Oratore

Abstract

This contribution argues that speech technologies, specifically speaker verification, speech recognition and speech synthesis, need to model the effects of speaker state (attitudes and emotions) in order to increase their quality and acceptance. Experimental research on the effects of stress and emotion on voice quality and prosody is reviewed and linked to basic dimensions of speech communication. It is concluded that the current state of the art in this area can provide speech technologists with important leads for modeling affective speaker states. We argue that there is a strong need for increased collaboration between speech engineers and speech scientists from other disciplines.

1. Introduction

Les développements récents dans le domaine du traitement des signaux ont permis l'essor d'un grand nombre de technologies de la parole dont beaucoup sont actuellement utilisées, voire sur le point d'être commercialisées. Cependant, leur fiabilité et leur usage sont fortement limités par le fait que la parole est affectée par des changements transitoires de l'état du locuteur, changements liés à son état émotionnel, à ses attitudes, à sa santé ou pouvant se produire sous l'impact de stress cognitif ou émotionnel.

Quoi qu'il soit compréhensible que les chercheurs et les ingénieurs en technologie du langage aient donné la priorité au développement d'algorithmes de base pour l'analyse et la synthèse de la parole, nous soutiendrons que continuer à négliger les facteurs affectifs mènera l'utilisateur à refuser cette technologie et privera cette dernière d'une entrée dans le marché à laquelle elle pourrait potentiellement prétendre.

La synthèse de la parole fournit un exemple parlant des effets d'une négligence des aspects pragmatiques de la parole, en particulier l'influence des attitudes et des émotions. En effet, bien qu'une technologie de synthèse raisonnablement accessible existe depuis un certain nombre d'années, ce qu'il en ressort actuellement est fréquemment jugé comme inacceptable pour la plupart des applications dans le monde réel. La raison est que la majorité des systèmes TTS (*text-to-speech*), entre autres, sont limités dans la manière dont ils prennent en compte les caractéristiques prosodiques de base pour différents types de phrases, comme les interrogations. Ils échouent même à fournir d'importants indicateurs stylistiques et affectifs tels que la politesse, l'intérêt, le plaisir ou, à d'autres fins, la détermination ou l'insistance. Nous croyons qu'il est essentiel que les ingénieurs travaillant sur la synthèse deviennent plus conscients de l'importance des indices pragmatiques et émotionnels et tentent de construire des outils appropriés pour communiquer cette signification pragmatique et affective.

Les utilisateurs du monde réel demanderont d'ailleurs eux-mêmes que les techniques de synthèse de la parole fournissent de tels indices avant d'accepter cette technologie à grande échelle. Evidemment, le développement d'instruments adéquats et de systèmes de règles exigera un regard scientifique sur les propriétés prosodiques et paralinguistiques des actes de la parole, des émotions et des attitudes du locuteur. Malheureusement, ce terrain de recherche a été plutôt négligé dans le passé par les ingénieurs de la parole, les linguistes et les phonéticiens, tout comme par les psychologues.

La vérification automatique du locuteur est un autre exemple de technologie de la parole qui pourrait être utilement appliquée dans de nombreux secteurs économiques. L'automatisation progressivement croissante de beaucoup d'activités dans le domaine du service, spécialement l'accès télécommandé à l'information, le paiement automatique par carte de crédit et la consultation bancaire par téléphone (*telephone banking)*, requièrent de plus en plus de moyens fiables pour assurer toute sécurité et limiter l'accès aux utilisateurs autorisés uniquement. Malgré des investissements conséquents et des tests en laboratoire prometteurs, la performance et l'acceptation de tels systèmes restent bien en-deça des attentes. Les systèmes de vérification du locuteur continuent à souffrir d'un nombre inacceptable d'erreurs, sous forme de faux rejets ou fausses identifications, et n'apparaissent pas suffisamment robustes pour un usage flexible dans les situations quotidiennes. En particulier, la performance de ces systèmes se dégrade avec le laps de temps entre l'entraînement et l'usage, un phénomène qui est probablement du, en majeure partie, à un changement d'état du locuteur [Dod98; Kar98].

Nous croyons que ces problèmes de performance et d'acceptation mentionnés ci-dessus sont partiellement dus au fait que beaucoup d'ingénieurs de la parole ou de scientifiques qui ont été particulièrement impliqués dans le développement de cette technologie ont négligé ou tout

au moins ont sous-estimé les effets puissants des variations transitoires de l'état organique du locuteur sur ses productions vocales et, en conséquence, sur les paramètres acoustiques utilisés dans les algorithmes de ces technologies du langage. Il est une croyance largement répandue dans la communauté scientifique: le fait d'augmenter davantage la sophistication des algorithmes statistiques permettant la reconnaissance et la comparaison des patterns résoudrait les problèmes posés par les variations internes au locuteur, variations produites par des changements attitudinaux ou émotionnels de son état et qui jusqu'à présent ont été traitées essentiellement au hasard.

Cependant, à mesure que le temps passe, il devient de plus en plus difficile de nier la nature asymptotique des progrès dans ce domaine et il est actuellement de moins en moins probable qu'une augmentation de la sophistication par une approche purement statistique mènera à des résultats persuasifs. L'importance des déficits de performance des systèmes de vérification du locuteur concernant la variation intra-locuteur n'a toutefois jamais encore été établie puisqu'il n'existe actuellement, à notre connaissance, aucune base de données avec des échantillons contenant des variations intra-locuteur systématiquement récoltées, qui pourraient être utilisés pour comparer les différents systèmes.

Les recherches antérieures ont largement démontré que les changements d'états du locuteur modifient les paramètres acoustiques de sa parole de façon spécifique et consistante [pour une revue de la littérature dans ce domaine, voir par exemple Fri77, Mur88, Pitt93; Sch81; Sch86; Sch95; Sch90]. Incorporer à la technologie de vérification du locuteur cette connaissance des dépendances à un état organique pourrait alors mener à des systèmes plus robustes, en permettant à la fois la sélection d'ensemble de caractéristiques acoustiques plus appropriées (traitant les algorithmes et éprouvant les procédures) et le développement de techniques de production de la parole élaborée de façon à minimiser les effets des changements transitoires de l'état physiologique du locuteur.

Une certaine partie de la communauté scientifique travaillant sur la parole, principalement les chercheurs orientés vers une approche multidisciplinaire de la technologie du langage, a insisté pendant longtemps sur le fait qu'une percée réelle dans la performance et l'acceptation des technologies de la parole ne peut être réalisée qu'en basant les développements ultérieurs sur une connaissance scientifique plus approfondie de la régularité et des lois des variations intra- et inter-locuteurs (voir, par exemple, les commentaires conclusifs de Fur94) ainsi que sur le rôle des indices pragmatiques dans la communication verbale. Un développement de notre compréhension des processus impliqués nécessite des travaux expérimentaux plus importants introduisant une manipulation systématique des états cognitifs, affectifs et attitudinaux du locuteur dans des populations distinctes de locuteurs. Ceci n'a jamais encore été systématiquement fait.

Notre groupe de recherche participe quelque temps aux efforts déployés pour étudier les effets des changements engendrés par différents types d'états physiologiques du locuteur sur les paramètres acoustiques de sa voix. Nous avons étudié les influences à court et long terme du stress, de l'émotion, des attitudes et des mensonges sur la voix et la parole, en utilisant des méthodes de traitement digital des signaux pour l'extraction de paramètres [Ban96; Lad85; Sch77; Sch84; Sch91; Tol86; Wal86]. Ceci a supposé le développement de techniques d'induction fiable et réaliste d'affects et de stress, basées sur des modèles théoriques testés empiriquement concernant les changements dans la voix suite à un changement affectif. Actuellement, nous poursuivons nos travaux en examinant les effets vocaux et acoustiques de l'induction d'événements émotionnels dans le cadre de jeu vidéo et autres interactions homme-ordinateur [Joh96]. Dans ce qui suit, nous soulignerons certaines questions et certains résultats que nous croyons pertinents pour une avancée dans la sophistication des technologies du langage.

Nous commencerons par exposer les principes de base sur la nature de la communication vocale qu'il s'agit de garder à l'esprit quoique travaillant sur des questions appliquées, plus moléculaires. Il est utile de débuter avec la communication animale, étant donné la continuité phylogénétique dans l'expression vocale des affects [Sch88].

2. Fonctions et déterminants multiples des vocalisations

Si l'expression vocale chez les animaux a d'abord été considérée comme un indicateur de leur état affectif ou motivationnel sous-jacent, les recherches plus récentes indiquent que la situation est plus complexe. Marler et ses collègues [Mar84], en étudiant les appels d'alarme chez une espèce de singe, ont trouvé que les appels ne sont pas seulement des indicateurs de leur peur mais qu'ils sont également liés spécifiquement aux types de prédateurs. Les cris d'alarme produits pour les léopards, les aigles ou les serpents, par exemple, présentent des sons, niveaux d'énergie et fréquences différents. C'est pourquoi, les auteurs ont rejeté l'hypothèse selon laquelle la communication animale serait limitée à n'indiquer que l'état émotionnel ou motivationnel de l'animal, en postulant que la plupart des appels des animaux ont un composant symbolique très fortement référenciel, partiellement appris. Le système d'alarme n'externaliserait pas seulement l'affect sous-jacent mais refléterait encore la classification du prédateur sur la base de processus cognitifs rudimentaires.

Mettre un accent exclusif soit sur l'expression de l'affect ou de la motivation, soit sur sa fonction symbolique, c'est omettre le fait que la plupart des signaux vocaux sont en fait pluri-fonctionnels. Le modèle " Organon " développé par Bühler [Büh34] peut être évoqué pour analyser les fonctions distinctes des affects vocaux. Dans ce modèle, un signe possède trois fonctions: il est un symbole, en représentant l'objet, l'événement ou le fait; il est un symptôme de l'état de l'utilisateur; il est enfin un appel ou un signal en ce qu'il tente de susciter une réponse du

récepteur. L'appel d'alarme du singe notamment comprend ces trois aspects : en tant que symbole de différents prédateurs, en tant que symptôme de l'état de peur de l'animal et en tant qu'appel aux autres membres de la troupe pour qu'ils s'enfuient. Par ailleurs, ces fonctions sont mutuellement inter-dépendantes. Si un appel se rapporte à un prédateur des airs, la réaction et le signal seront différents de ceux qui seraient faits en référence à un prédateur terrestre: dans le premier cas, l'émetteur et le récepteur tous deux devraient chercher une cachette sous des branches ou des buissons; dans le second cas, ils devraient s'activer physiologiquement et grimper à un arbre.

Outre ces multiples fonctions, nous devons considérer de multiples déterminants. Nous avons suggéré de distinguer les termes " effet push " et " effet pull ". Les effets " push " concernent les processus physiologiques, telle que la tension musculaire, qui " poussent " les vocalisations dans une certaine direction. Ils reflètent les changements dans les sous-systèmes physiologiques de l'organisme et ayant un effet direct sur les paramètres vocaux. Les effets " pull " ont trait, quant à eux, à des facteurs externes, telles les attentes de l'auditeur, qui " tirent " la vocalisation de l'affect vers un modèle acoustique particulier. Les facteurs " pull ", bien que médiatisés par des systèmes internes, sont basés sur l'extérieur. Ils opèrent vers une production de patterns ou de modèles acoustiques spécifiques, définis et valorisés socialement. Cette distinction est importante pour la compréhension des différences entre les productions vocales.

Ainsi, l'information dans la voix reflète aussi bien les modifications physiologiques, liées à l'état organique, et totalement involontaires qui affectent les systèmes de production du langage du locuteur (effet " push ") que l'adoption de styles de langage culturellement acceptés (effet " pull ") [Sch86]. Cette notion de changements involontaires et volontaires dans la production de la parole est cruciale pour une vérification du locuteur. Les modifications physiologiques à la base de la glotte sont susceptibles de produire des changements plutôt à long terme, supra-segmentaux du signal acoustique. Ceux-ci se manifesteront dans les paramètres acoustiques telles que la hauteur, l'intensité ou la distribution spectrale de l'énergie. Les changements dans l'articulation produisent des modifications plutôt à court terme et segmentales du signal acoustique mesurées par la fréquence et la précision des formants. Toutefois, ces modifications segmentales pourraient aussi susciter des changements à long terme lorsqu'elles se prolongent sur de longues périodes de la parole.

Ce concept d'effets " pull/push " en tant que deux types majeurs de déterminants des signaux vocaux nous paraît directement lié au modèle des multiples fonctions de Bühler. On pourrait prétendre que la fonction de symptôme, c'est-à-dire l'expression d'un état interne représente les effets " push ", alors que le symbole et l'appel seraient des aspects représentant les effets " pull ". Différents facteurs pourraient déterminer la nature de l'expression dans chaque cas. Et il pourrait y avoir un antagonisme entre effets " push " et " pull ", comme dans le cas par exemple où une excitation physiologique accrue " pousserait " la fréquence fondamentale à un niveau plus élevé, alors que les tentatives conscientes de contrôle la " tireraient " vers le bas, ce qui peut mener à la production de messages mixtes, voire contradictoires. S'il en est ainsi, il serait empiriquement important d'isoler ces deux déterminants. Recherche et théorisation futures dans ce domaine seront nécessaires pour différencier plus exactement entre ces multiples déterminants et multiples fonctions afin d'éviter de futiles controverses sur la " vraie nature " des vocalisations affectives.

3. Evaluation empirique des déterminants multiples

Le type de déterminant devrait avoir un effet important sur le codage, c'est-à-dire sur la relation entre le référent sous-jacent et les caractéristiques du signal. Dans une condition " push ", la tension des muscles augmenterait sous l'effet du stress en produisant une augmentation de la fréquence fondamentale de la voix (Fo). On s'attend à une covariation directe entre la quantité d'augmentation de tension musculaire, mesurée par électromyographie, et l'augmentation de la fréquence fondamentale. Dans ce modèle, la covariation attendue est continue (linéaire ou non-linéaire) entre les deux classes de variables.

Une alternative à ce modèle serait celui que nous appellerons le modèle configuration. Ce modèle est plus " linguistique " que le modèle covariation, lui-même plus psychologique par nature. Le modèle configuration soutient que pour réaliser un certain effet sur l'auditeur, l'individu utilise une combinaison particulière d'intonation, d'accent, de mots et de structures syntaxiques, par exemple une élévation du contour de l'intonation dans une interrogation et une diminution de cette intonation pour une question appelant une réponse de type oui/non. Il n'y a pas de dimensions variables ni continues dans les effets du modèle de configuration. Certaines classes de phénomènes doivent se coproduire pour susciter un effet. En terme d'effets " push " et " pull ", " push " est susceptible de suivre les règles du modèle de covariation, alors que " pull " suivrait plutôt les règles du modèle de configuration. Afin de mieux comprendre comment les processus de communication suivent soit le modèle de covariation soit celui de configuration, nous avons besoin de récolter davantage de connaissances sur les déterminants.

Scherer, Ladd et Silvermann [Sch84] ont mené deux études pour distinguer covariation et configuration. La première a utilisé un corpus de questions tirées d'une vaste étude sur les interactions entre fonctionnaires et citoyens. Celles-ci étaient homogènes dans leur structure mais variaient en terme de style pragmatique. Certaines questions étaient clairement des reproches, formulés comme des interrogations, alors que d'autres étaient des questions d'information purement factuelle. Trois techniques de filtrage ont été utilisées *a) low pass filtering, b)"random splicing" et c) "reversing"* (voir Sch85 pour une liste comparative des indices acoustiques

retenus par les techniques respectives). Les résultats démontrent que même lorsque le contenu des questions est rendu inintelligible, une grande part de la signification affective reste dans le signal acoustique. Le modèle de covariation se trouve confirmé dans son affirmation que les indices vocaux non-verbaux convoient l'affect de façon directe et indépendante du contexte.

Néanmoins, ces résultats n'apparaissent que pour les conditions de masquage dans lesquelles les paramètres de la qualité de la voix sont audibles, c'est-à-dire dans les techniques de *random splicing* et *reversing*. Dans ces deux conditions, le contour d'intonation de la phrase est perdu ou détruit. Ces caractéristiques ne joueraient donc pas un rôle essentiel dans la communication de la signification affective.

Evidemment, il y a des preuves empiriques contradictoires et contre-intuitives montrant que cette information est pertinente pour la communication de l'affect. Serait-ce que l'intonation suit les règles du modèle de configuration plutôt que celles du modèle de covariation ? Afin d'étudier cette question, avec le matériel de cette étude, nous avons divisé les questions en interrogations et en questions appelant des réponse de type oui/non et classifié les contours d'intonation en chute finale ou en augmentation finale. Les résultats montrent qu'il y a une forte interaction entre type de question et courbe d'intonation. Donc, on constate que certains paramètres acoustiques, telle que la qualité de la voix, opèrent selon les règles du modèle de covariation alors que d'autres, tel que le contour d'intonation, sont utilisés selon les règles du modèle de configuration.

Cet argument gagne davantage de sens si on le considère en terme d'approche psychobiologique de la communication. En effet, les paramètres qui montrent un degré remarquable de continuité phylogénétique, comme la nature différentielle de la phonation qui mène à des qualités de voix différentes, pourraient être plus proches du modèle de covariation directe avec les états physiologiques. A l'inverse, les paramètres qui ont été " domestiqués " avec le système de langage, telle que l'intonation, suivraient un modèle de configuration.

Afin de tester davantage ces notions, nous avons utilisé des techniques de synthèse et resynthétisation digitale de la parole pour évaluer l'influence d'indices prosodiques spécifiques et de la qualité de la voix sur l'impression de l'auditeur. Une telle approche évite évidemment le désavantage d'utiliser un corpus naturel, puisqu'elle permet un contrôle expérimental plus étroit des variables étudiées qui peuvent être manipulées indépendamment l'une de l'autre, tout en maintenant les autres paramètres acoustiques constants. Dans une série de travaux, nous avons employé cette technique pour varier systématiquement le contour d'intonation, la fréquence fondamentale (Fo), l'intensité, le rythme, l'accent, la structure et d'autres paramètres. [Ber88; Lad85; Tol88].

Trois résultats essentiels ont été mis en évidence: d'abord, nous n'avons pas trouvé d'effet d'interaction, ce qui suggère que les variables acoustiques que nous avons étudiées fonctionnent largement indépendamment l'une de l'autre. Deuxièmement, dans ces études où nous avions plusieurs locuteurs et plusieurs phrases, nous n'avons en réalité trouvé aucune interaction entre ces facteurs et les paramètres acoustiques manipulés. Ceci nous encourage à penser que les effets peuvent être généralisés à une grande variété de locuteurs et de phrases. Enfin, la fréquence fondamentale se présente comme celle qui a de loin le plus puissant effet sur le juge, particulièrement sur ses attributions de l'excitation. De plus, ces attributions semblaient être une fonction continue des changements dans la variabilité de la fréquence fondamentale, puisque les augmentations d'excitation rapportées étaient linéaires avec les augmentations de la fréquence fondamentale. Les résultats pour les contours d'intonation et la qualité de la voix sont complexes et demandent encore à être précisés par des travaux ultérieurs. En ce qui concerne les contours d'intonation, la difficulté pourrait être due en partie au rôle important joué par le modèle de configuration pour cette variable; en conséquence, nous considérons que la distinction entre règles de configuration et règles de covariation pourrait être très utile pour la compréhension de la communication de l'affect par la parole et il nous apparaît utile de continuer ce type de recherches avec l'aide des techniques modernes de manipulation des signaux digitaux.

4. Les profils acoustiques de l'émotion

Les travaux mentionnés ci-dessus ont trait à des états affectifs d'une intensité relativement basse, proches de ceux que nous rencontrons dans nos interactions sociales normales. Les études sur l'encodage, c'est-à-dire concernant l'effet des changements physiologiques dans l'état du locuteur sur ses productions vocales, ont été menées en induisant expérimentalement un état émotionnel spécifique ou en utilisant des prestations d'acteurs simulant un état émotionnel afin d'établir, par des méthodes de traitement digital, les conséquences de ces états affectifs sur la voix et la parole [Sch84; Sch91; Wal86]. Pittam et Scherer [Pitt93] ont résumé l'état de la littérature à ce jour comme suit :

COLERE : elle semble être, généralement, caractérisée par une augmentation de la moyenne de la F0, de la moyenne de l'énergie, de l'énergie haute fréquence ainsi que par des contours de la F0 descendante. La vitesse d'articulation habituellement augmente. Certaines études, qui ont mesuré la colère " chaude " ou rage (la plupart des études ne définissent pas si elles étudient la rage ou la colère " froide " plutôt synonyme d'irritation) ont également montré une augmentation de la variabilité de la F0 et du champ de la F0 à travers les phrases encodées. Les études qui n'ont pas trouvé de telles caractéristiques pourraient avoir mesuré en fait la colère froide.

PEUR : il y a un accord important sur les paramètres acoustiques associés à la peur. On s'attend à ce qu'un niveau d'excitation soit associé à cette émotion et ceci semble appuyé par l'augmentation de la moyenne de la F0, du champ de la F0 et de l'énergie haute fréquence généralement observée. La vitesse d'articulation est plus élevée. Une augmentation dans la moyenne de la F0 a également été trouvée dans des formes plus douces de

l'émotion, comme dans l'inquiétude ou l'anxiété.

TRISTESSE : comme pour la peur, on rencontre une bonne convergence entre les études qui ont pris en compte cette émotion. Une diminution de la F0, de l'étendue de la F0 et de la moyenne de l'énergie est en principe observée, de même qu'une tendance des contours de la F0 descendante. L'énergie haute fréquence et la vitesse d'articulation diminuent. La plupart des recherches ont toutefois étudié des formes plus tranquilles de cette émotion plutôt que des variantes très aiguës comme le désespoir. Ce dernier pourrait être caractérisé par une augmentation de la F0 et de l'énergie.

JOIE : c'est l'une des rares émotions positives étudiées, le plus souvent sous sa forme " forte " comme l'allégresse plutôt que sous ses types atténués comme le plaisir ou le contentement. Consistant avec un niveau élevé d'excitation auquel on pouvait s'attendre, nous avons trouvé une forte convergence des recherches sur une augmentation de la moyenne de la F0, de l'étendue de la F0 et de la variabilité de la F0, de même que de la moyenne d'énergie. Certains travaux ont mis de plus en évidence une augmentation de l'énergie haute fréquence et de la vitesse d'articulation.

DEGOUT: les résultats pour le dégoût tendent à être inconsistants au travers des études. Le peu d'entre elles ayant inclus cette émotion varient dans leur procédure d'encodage pour la mesure du dégoût, s'appuyant tantôt sur des films désagréables, tantôt sur des prestations d'acteurs simulant l'émotion. Les études utilisant des films ont relevé une augmentation de la moyenne de la F0, alors que celles utilisant des simulations d'acteurs ont indiqué l'inverse, à savoir une diminution de la F0. Cette inconsistance trouve écho dans la littérature sur le décodage.

Les travaux empiriques ne se sont pas limités à l'étude d'émotions primaires mais se sont également progressivement intéressés aux effets du stress ou encore aux états affectifs pathologiques, telle la dépression ou la schizophrénie, qui semblent également s'accompagner de caractéristiques vocales spécifiques.

STRESS: une déviation de la F0 à partir de sa ligne de base paraît indiquer les changements de magnitude d'un stress émotionnel [Kur76; Sim80; Gri87]. Plus spécifiquement, un stress émotionnel ou des conditions de charges mentales intenses mènent à une élévation des valeurs de la F0. Des effets similaires ont été mentionnés concernant l'amplitude et la durée des phrases parlées: chez des sujets vivant un stress majeur, l'amplitude moyenne de l'enveloppe spectrale est généralement plus grande alors que la moyenne de durée des phrases est plus basse [Alp63; Gri87]. Au contraire, Tolkmitt et Scherer [Tol86] ont trouvé que le parterre de la F0 est plus élevé lors de stress ou cognitif ou émotionnel chez des sujets très anxieux ou répresseurs, alors que chez des sujets peu anxieux ce changement de F0 est opposé. De plus, et indépendamment du type de stress, une modification de la précision des deux premiers formants a été observée, la direction de ces variations dépendrait toutefois des stratégies de coping utilisées par le locuteur.

DEPRESSION : les résultats les plus consistants concernent l'effet de la dépression sur l'intensité de la voix. Une étude menée par Ellgring et Scherer [Ell96], a pu confirmer un niveau de F0 élevé dans les phases dépressives aigues, amélioré après traitement (comme constaté par des recherches antérieures, p.e. Ekm76; Tol82), mais seulement pour des femmes dépressives, pas pour les hommes. Par contre, ces auteurs trouvent un ralentissement du débit chez les deux sexes.

Même si ces résultats semblent indiquer une différenciation acoustique relativement claire entre les émotions fondamentales, on ne peut cependant pas exclure que la majeure partie des différences soit due au simple facteur excitation – excitation sympathique élevée, qui est caractéristique de plusieurs émotions et qui serait responsable de l'augmentation de la F0, de l'énergie et de l'énergie spectrale haute fréquence. La question de savoir si l'expression vocale indique seulement une excitation sympathique, plutôt que des différences qualitatives entre émotions (comme trouvé dans le cas des expressions faciales prototypiques) a été une autre préoccupation importante dans le domaine.

Une étude récemment menée par notre groupe de recherche permet une avancée concernant la question [Ban96]. Douze acteurs professionnels ont été priés de jouer 14 émotions en variant intensité et valence (qualité). Au total, 224 prestations différentes ont été recueillies, et ont été présentées à des juges auxquels on demandait de décoder ou d'inférer la catégorie émotionnelle renvoyée par l'émetteur. Les résultats répliquent les recherches antérieures démontrant la capacité des juges à inférer des émotions exprimées vocalement avec une précision supérieure à celle qui pourrait être attribuée à la chance. Le tableau 1 présente les différences dans la précision de reconnaissance pour 14 émotions.

Tableau 1: Classification des représentations d'émotions vocales par des juges, *jack-knifing*, et analyse discriminante en pourcentage

	Emotions représentées														
Emotions reconnues	CCh	CFr	PPan	Anx	Dés	Tris	Allég	Bonh	Int	Ennu	Hon	Fier	Dég	Mép	*Sum*
Colère chaude	**78**	17	10		6		14							2	127
	69		13		25		6		6						119
	75		13		13		13		6						120
Colère froide	10	**34**	2	7	5	1	5	1	3	2	2	5	10	10	97
		44		7				6	13			13	19	6	108
	13	**50**	6						13	6		13	13		114
Peur panique			**36**	13	9	1	7				1		1		68
	6	19	**31**		13		6								75
	6	13	**50**		6		13								88
Anxiété			27	**42**	18	2	5	1	1		15	2	10	2	125
				60	6	6		13	6	19	25	25		6	166
				53	6	25		19		6	19	13			141
Désespoir			21	7	**47**	21	16	1		1	4		10		128
	6	13	19		**38**		13								89
		13	13		**50**		13								89
Tristesse				5	8	**52**		3		13	19	2	14	8	124
			6			**44**				19	13	6	6	6	100
		6				**63**					13	6	6	6	100
Allégresse		1			1		**38**	2				4			46
	13		19		19		**69**								120
	6		6		19		**63**								94
Bonheur		2		4	1	3	1	**52**	8	1	8	23	5		108
				20		13		**44**			19	13	6	6	121
		6						**50**	6		6			13	81
Intérêt		7	1	7		1	4	18	**75**	1	13	12	2	2	143
		13	6	7				13	**56**		6	19	6		126
				7				6	**50**		13	13		6	95
Ennui		4		1		5		4	1	**76**	4	2	2	4	103
		6				19				**38**	6		6	13	88
		6		13	6					**56**	6		6	13	106
Honte			1	5	2	9	1	3	1	1	**22**	2	10	2	59
								13			**13**		6		31
				7				6		6	**31**		6		56
Fierté	1	15	1	4		2	2	17	10	1	8	**43**	7	6	117
	6		6	7		6	6	6		6		**13**	25	19	100
		6	6	7		6		13	13		6	**56**	19		132
Dégoût	1	2	1	2	2		2		1		1		**15**	5	32
		6				6		6	13	6	13	6	**13**	6	75
			6	13		6			6	6			**50**	25	112
Mépris	11	18	1	4	3	3	7	1	1	6	4	6	15	**60**	140
						6			6	13	6	6	13	**38**	88
								6		19	6	6		**38**	75

Note: Les résultats de l'étude sur la reconnaissance sont présentés dans la première ligne de chaque cellule, les résultats du *jack-knifing* et de l'analyse discriminante dans la seconde et troisième ligne respectivement. Les cellules vides représentent les valeurs zéro.

Les différences mesurées de façon consistante dans la " reconnaissabilité " de diverses émotions furent également répliquées. Toutes les prestations des acteurs furent soumises à des analyses acoustiques digitales de façon à obtenir des profils des paramètres vocaux pour chaque émotion, en utilisant un grand nombre de variables acoustiques. Le tableau 2 présente les profils acoustiques pour les émotions les plus intéressantes.

Les données indiquent que les paramètres vocaux non seulement indexent le degré d'intensité typique pour les diverses émotions mais différencient également la valence ou les aspects qualitatifs. Elles suggèrent de plus qu'avec davantage de finesse dans la mesure acoustique, il pourrait être possible de déterminer des profils acoustiques stables pour la plupart des émotions, pourvu que les états émotionnels différenciés soient utilisés de façon appropriée. Des analyses discriminantes et *jack-knifing* ont été conduites afin de déterminer avec quel degré de réussite les 14 émotions pouvaient être différenciées sur la base des paramètres vocaux mesurés. Les résultats montrent de façon remarquable que les réponses correctes et les patterns de confusion reflètent étroitement ceux donnés par des juges. Les profils acoustiques de la tristesse et de la honte (Figure 1) illustrent comment deux émotions exprimées avec des charactéristiques acoustiques semblables sont hautement confondues, autant par des juges humains que par des classifications informatisées. On a pu également trouver qu'une bonne part de la variance dans les inférences des juges pouvait être prédite sur la base des mesures acoustiques en permettant une évaluation plus détaillée des indices acoustiques que les auditeurs utilisent dans l'inférence de l'émotion à partir de la voix.

Tableau 2: Profils vocaux de 14 émotions: moyennes et déviations standards de 29 paramètres acoustiques résiduels transformés en scores z (Sexe de l'Acteur et Identité de l'Acteur partialisés).

	CCh	*CFr*	*PPan*	*Anx*	*Dés*	*Tris*	*Allég*	*Bonh*	*Int*	*Ennu*	*Hon*	*Fier*	*Dég*	*Mép*
MF0	1.13	.16	1.23	-.58	.99	-.32	1.24	-.64	-.17	-.80	-.49	-.46	-.29	-1.03
P25F0	.92	.15	1.39	-.28	1.15	-.52	1.21	-.62	-.14	-.83	-.64	-.51	-.37	-.93
P75F0	1.13	.05	.91	-.83	.73	-.08	1.20	-.52	-.32	-.69	-.41	-.37	.00	-.85
SdF0	.50	-.10	-.63	-.86	-.73	.43	.21	.14	-.26	.07	.42	.07	.33	.35
MElog	1.19	.52	.84	-.37	1.00	-1.16	1.05	-.48	.19	-.54	-1.14	-.13	-.51	-.48
Dur Art	-.31	-.14	-.58	-.35	.32	1.04	.12	-.49	-.66	.70	.32	-.22	.08	.15
DurVo	-.45	.15	-.47	-.38	.07	1.25	-.34	-.45	-.42	.94	.20	-.06	.01	-.06
HammI	1.13	.29	.27	-.33	.90	-.43	.58	-.43	-.03	-.40	-.49	-.26	-.46	-.37
DO1000	-1.17	-.51	-.45	.16	-.72	1.32	-.66	.15	-.23	.70	.89	.04	.45	.05
PE500	-.55	-.58	-.12	.15	-.51	1.23	-.29	.32	-.30	.27	.51	-.09	-.17	.12
PE1000	-1.34	-.52	-.28	.53	-.59	.90	-.05	.39	.11	.44	.03	.35	-.11	.17
v-0.2K	-.43	-.40	-.33	.12	-.69	.66	-.34	.07	-.24	.85	.48	-.12	-.11	.49
v-0.3K	-.59	-.37	-.19	.89	-.33	.81	-.57	.23	-.27	.21	.59	.02	-.06	-.31
v-0.5K	-.13	-.36	.13	-.54	-.16	.85	.20	.25	-.10	-.19	.05	-.10	-.14	.19
v-0.6K	-.31	-.33	-.28	.62	-.13	-.05	.11	.11	.09	.16	-.18	.09	-.06	.19
v-0.8K	-.08	.62	-.14	-.18	.20	-.46	-.18	.00	.10	.13	-.21	.18	.30	-.29
v-1K	-.13	.18	.42	-.25	.18	-.66	.56	-.28	.49	-.32	-.46	.29	-.09	.06
v1-1.6K	1.46	.63	.30	-.44	.64	-.73	.09	-.56	.04	-.55	-.35	-.13	-.22	-.22
v1.6-5K	.86	.27	.17	-.44	.36	-.84	.04	-.07	-.21	-.26	.28	-.45	.33	-.05
v5-8K	-.33	-.12	.11	-.18	-.08	.38	-.31	-.21	-.24	.48	.18	-.23	.72	-.19
uv-0.25K	-.65	-.18	-.30	.33	-.56	1.04	-.65	.26	.39	.06	.46	.13	-.32	.00
uv-0.4K	.04	.52	.39	-.25	-.02	.14	-.30	.38	.10	-.36	-.31	-.13	.16	-.36
uv-0.5K	.75	-.17	.47	-.07	.01	.33	.14	-.15	.09	-.49	-.06	-.01	-.46	-.39
uv0.5-1K	-.12	.15	.27	-.23	.31	-.06	.05	.34	.02	-.44	-.34	.50	-.03	-.44
uv1-1.6K	.49	-.02	.14	-.08	.62	-.46	1.20	-.29	-.41	-.12	-.45	-.03	-.36	-.23
uv-2.5K	.62	.15	.04	-.04	.75	-.79	.43	.16	-.42	-.29	-.23	-.28	-.25	.14
uv2.5-4K	-.19	-.09	-.45	.28	-.72	-.11	-.61	-.06	.56	.19	.59	.17	.12	.32
uv4-5K	-.57	-.11	-.36	.17	-.52	.50	-.61	-.08	.15	.64	.22	-.22	.42	.39
uv5-8K	-.49	-.30	.11	-.04	-.40	.80	-.40	-.31	-.15	.51	.35	-.27	.37	.21

Légende: Fréquence fondamentale: MF0 = Moyenne, SdF0 = déviation standard, P25F0 = 25th pourcentile, P75F0 = 75th pourcentile; Energie: MElog = Moyenne; Taux de parole: DurArt = durée des périodes d'articulation, DurVo = durée des périodes de voix; Spectre moyen vocalisé: v-0.2K = 125-200 Hz, v-0.3K = 200-300 Hz, v-0.5K = 300-500 Hz, v-0.6K = 500-600 Hz, v-0.8K = 600-800 Hz, v-1K = 800-1000 Hz, v1-1.6K = 1000-1600 Hz, v1.6-5K = 1600-5000 Hz, v5-8K = 5000-8000 Hz; LMHamm = Indice Hammarberg; DO1000 = pente d'énergie spectrale au-dessus de 1000 Hz; PE500 = proportion d'énergie vocale jusqu'à 500 Hz; PE1000 = proportion d'énergie vocale jusqu'à 1000 Hz.; Spectre moyen non vocalisé: uv-0.25K = 125-250 Hz, uv-0.4K = 250-400 Hz, uv-0.5K = 400-500 Hz, uv0.5-1K = 500-1000 Hz, uv1-1.6K = 1000-1600 Hz, uv-2.5K = 1600-2500 Hz, uv2.5-4K = 2500-4000 Hz, uv2.5-4K = 4000-5000 Hz, uv5-8K = 5000-8000 Hz.

Les caractéristiques acoustiques d'une vocalisation émotionnelle pourraient refléter le pattern complet des processus d'évaluation cognitive qui a déterminé l'état émotionnel chez le locuteur. Cette évaluation des

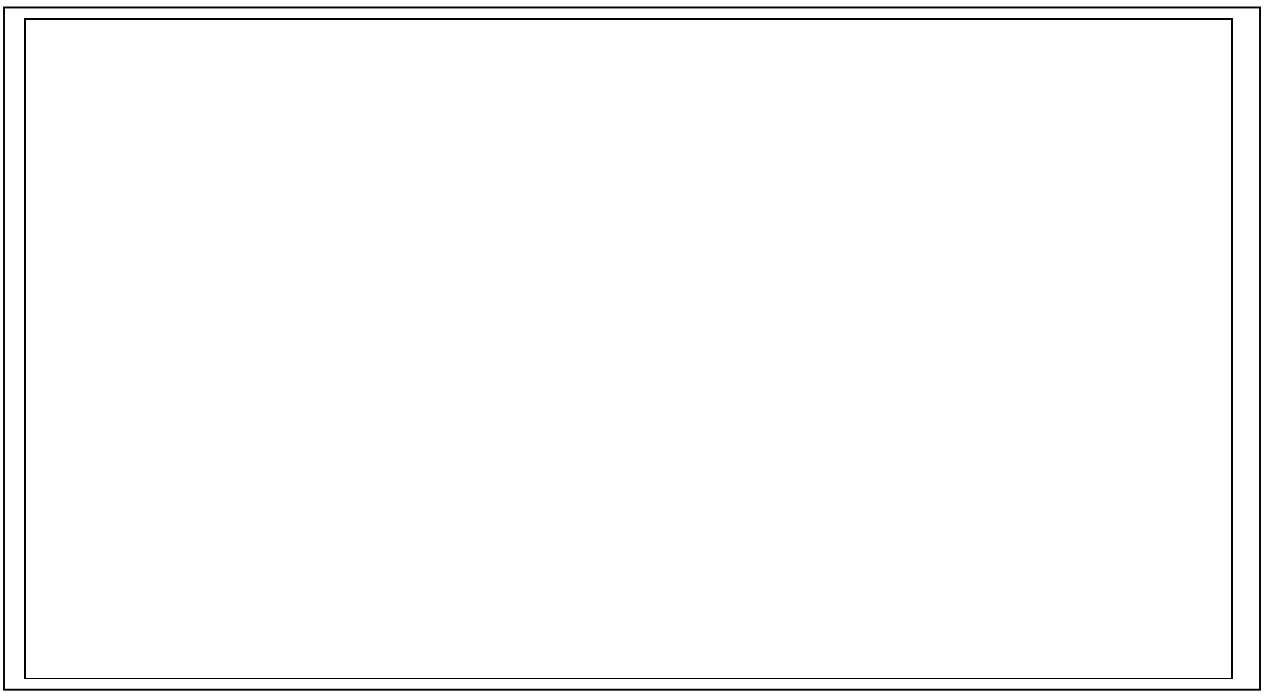

Figure 1 : Profils acoustiques de la tristesse et de la honte. Légende : voir tableau 2.

antécédents de l'émotion pourrait permettre à l'auditeur de reconstruire les caractéristiques majeures de l'événement et de ses effets sur le locuteur. Afin d'expliquer ce postulat, nous reprendrons certaines théorisations récentes sur l'émotion. Beaucoup de théoriciens du domaine de la psychologie de l'émotion semblent convaincus que la plupart des émotions humaines sont précédées par une évaluation cognitive de l'événement et de la situation, bien que les types de processus cognitifs puissent être relativement de bas niveau et non-conscients. Il est possible d'élaborer des prédictions sur comment on s'attendrait à ce que les caractéristiques phonatoires principales varient selon les critères d'évaluation des antécédents de l'émotion. Les données obtenues à partir de prestations d'acteurs furent utilisées pour tester ces prédictions théoriques sur le patterning vocal, basé sur le modèle des processus composant de l'émotion [Sch74]. Bien que la plupart des hypothèses soient supportées, certaines ont besoin d'être révisées sur la base des preuves empiriques accumulées [Ban96] .

5. Conclusion

La revue précédente de recherche sur l'effet de l'émotion sur la voix et la parole montre qu'une connaissance suffisante a été acquise pour commencer à essayer de transférer une partie de ces questions appliquées dans la technologie de la parole, en particulier la vérification du locuteur et la reconnaissance et synthèse de la parole. Evidemment, comme toujours, des recherches supplémentaires sont nécessaires, surtout depuis que la plupart de ce qui existe dans la littérature est limité aux langues anglaises et allemandes, et accompli par un nombre limité d'équipes de recherche. Toutefois, une grande part de cette recherche future pourrait être menée dans des cadres appliqués, abandonnant ainsi la distinction vaine entre recherche de base et recherche appliquée. Plus que dans beaucoup d'autres domaines, la recherche sur les effets du stress et de l'émotion sur la parole peut être menée avec profit dans le cadre d'un agenda de recherche ayant des buts appliqués clairement définis. Un exemple est un projet en cours sur la vérification du locuteur, menée par notre groupe (partiellement en collaboration avec d'autres laboratoires européens dans le projet VeriVox; Kar98]. Le but est d'entraîner des algorithmes de vérification du locuteur sur un assortiment plus large d'énonciations d'entraînement, contenant des échantillons obtenus sous différentes conditions émotionnelles. On s'attend à ce que l'élaboration de modèles qui définissent l'espace du locuteur quant à ses extensions potentielles dues aux effets émotionnels et attitudinaux, aidera à augmenter la robustesse (solidité) générale des algorithmes. En poursuivant ce but appliqué, l'étude produit également d'importants résultats de recherche de base sur la nature des signaux acoustiques qui différencient les états émotionnels et leur variabilité à travers les locuteurs, ainsi que beaucoup d'autres éclaircissements importantes.

L' exigence majeure pour une synergie prospère dans les efforts des ingénieurs tentant de perfectionner différents outils dans la technologie du langage, et des chercheurs provenant de disciplines contribuant à la science de la parole (telles que la phonétique et la psychologie), est l'abandon de la confiance exclusive dans la fiction selon laquelle les algorithmes statistiques améliorés s'occuperont des différences individuelles et des changements dans l'état du locuteur. Nous espérons que cette contribution puisse aider à fournir d'autres évidences que ceci consiste à prendre ses désirs pour des réalités plutôt que d'avoir des attentes réalistes. Comme il est montré au début de cette contribution, l'expérience pratique avec cette approche a été moins concluante jusqu'à maintenant, les critères de performance et d'acceptation étant quelque peu mal ciblés**.** De plus, nous avons tenté de montrer dans le reste de l'article, que des raisons théoriques suggèrent la nécessité de construire une compétence obtenue par une meilleure compréhension des mécanismes sous-jacents dans les algorithmes. Alors qu'une approche basée exclusivement sur les algorithmes d'ingénierie peut fonctionner dans certains domaines, il semble peu probable que la parole, en tant que système complexe de signaux combinant à la fois les systèmes de signification configurationnels et covariationnels, soit l'un d'entre eux.

De toute évidence, ce serait un grand ordre pour les ingénieurs de la parole de creuser dans les complexités de ce qui est déjà un domaine interdisciplinaire exigeant une compétence psychologique, physiologique et phonétique. Quoi qu'il en soit, il y a des exemples de collaboration fructueuse entre les ingénieurs et les scientifiques de la parole de différents milieux. Il nous semble que c'est ici la direction à prendre et nous désirons clore cette contribution par un appel à une collaboration accrue entre les différents groupes de chercheurs et de praticiens dans la communauté de la technologie de la parole.

Bibliographie

[Ban96] Banse, R. & Scherer, K. R. (1996), Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70, 614-636.

[Ber88] Bergmann, G., Goldbeck, T., & Scherer, K. R. (1988). Emotionale Eindruckswirkung von prosodischen Sprechmerkmalen. *Zeitschrift für experimentelle und angewandte Psychologie*, 35, 167-200.

[Büh34] Bühler, K. (1934). *Sprachttheorie*, Jena: Fischer (new edition 1984).

[Dod98] Doddington, G. R. (1998). Speaker Recognition Evaluation Methodology. Proceedings of IEEE/ESCA RLA2C Workshop, Avignon.

[Ekm76] Ekman, P., Friesen, W. V., & Scherer, K. R. (1976). Body movement and voice pitch in deceptive interaction. *Semiotica*, 16, 23-27.

[Ell96] Ellgring, H. & Scherer, K. R. (1996). Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior*, 20 (2), 83-110.

[Fri77] Frick, R. W. (1977). Prophylaxia in psychosomatic gynecology, as seen by a psychologist. *Psychotherapie und medizinische*

Psychologie, 27 (2), 53-57.

[Fur94] Furui, S. (1994). An overview of speaker recognition technology, Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, 1-9.

[Gri87] Griffin, G. R. & Williams, C. E. (1987). The Effects of Different Levels of Task Complexity on Three Vocal Measures. *Aviation, Space, and Environmental Medicine*, 58, 1165-1170.

[Joh96] Johnstone (1996). Emotional speech elicited using computer games, Proceedings of the 4th International Conference on Spoken Language Processing.

[Kar98] Karlsson, I., Banziger, T., Dankovicová, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., & Scherer, K. R. (1998). Speaker verification with elicited speaking-styles in the Verivox project. Proceedings of IEEE/ESCA Workshop on Speaker Recognition and its Commercial and Forensic Applications, Avignon.

[Kur76] Kuroda, I., Fujiwara, O., Okamura, N. & Utsuki, N., (1976). Method for Determining Pilot Stress Through Analysis of Voice Communication. *Aviation, Space, and Environmental Medicine*, 47, 528-533.

[Lad85] 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.

[Mar84] Marler, P. (1984). Animal communication: Affect or Cognition ? In K. R. Scherer & P. Eckman (Eds.), *Approaches to emotion* (pp. 345-368). Hillsdale, N. J. : Erlbaum

[Murr88] Murray, I. R., Arnott, J. L., & Newell, A. F. (1988). HAMLET-Simulating emotion in synthetic speech. Proceedings of Speech '88, The 7th FASE Symposium, Edinburgh, 4, 1217-1223.

[Pitt93] Pittam, J. & Scherer, K.R. (1993). Vocal Expression and Communication of Emotion. In M. Lewis and J. M. Haviland (Eds.) *Handbook of Emotions*. New York:Guilford Press.

[Sch74] Scherer, K. R. (1974a). Voice quality analysis of American and German speakers. *Journal of Psycholinguistic Research*, 3, 281-290.

[Sch77] Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilisation in emotion attribution from auditory stimuli. *Motivation and Emotion*, 1, 331-346.

[Sch81] Scherer, K. R. (1981). Speech and emotional states. In J. Darby (Ed.), *Speech evaluation in psychiatry*, 189-220. New York: Grune & Stratton.

[Sch84] 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.

[Sch85] 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.

[Sch86] Scherer, K. R. (1986). Vocal Affect Expression: A review and a Model for Future Research. *Psychological Bulletin*, 99, 143-165.

[Sch88] Scherer, K. R., Kappas, A. (1988). Primate Vocal Expression of Affective State. In D. Todt, P. Goedeking, & D. Symmes (Eds.), *Primate Vocal Communication* , 171-194. Berlin: Springer-Verlag.

[Sch90] Scherer, K. R., & Wallbott, H. G. (1990). Ausdruck von Emotionen. In K. R. Scherer (Ed.) *Enzyklopaedie der Psychologie*. Band C/IV/3 Psychologie der Emotion, 345-422. Göttingen: Hogrefe.

[Sch91] Scherer, K.R., Banse, R., Wallbott, H.G., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, 15, 123-148.

[Sch95] Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of Voice*, 9(3), 235-248.

[Sim80] Simonov, P. V., Frolov, M. V. & Ivanov, E. A., (1980) Psychophysiological Monitoring of Operators Emotional Stress. *Aviation and Astronautics. Aviation, Space, and Environmental Medicine* 51, 46-50.

[Tol82] Tolkmitt, F., Helfrich, H., Standke, R., & Scherer, K. R. (1982). Vocal indicators of psychiatric treatment effects in depressives and schizophrenics. *Journal of Communication Disorders*, 15, 209-222.

[Tol86] Tolkmitt, F. J. & Scherer, K. R. (1986). Effect of Experimentally Induced Stress on Vocal Parameters. *Journal of Experimental Psychology: Human Perception and Performance*, 12, 302-313.

[Tol88] Tolkmitt, F., Bergmann, T., Goldbeck, T, & Scherer, K. R. (1988). Experimental studies on vocal communication. In K. R. Scherer (Ed.), *Facets of emotion: Recent research*, 119-138. Hillsdale, NJ: Erlbaum.

[Wal86] Wallbott, H. G., & Scherer, K. R. (1986). Cues and channels in emotion recognition. *Journal of Personality and Social Psychology*, 51, 690-699.