

CARACTÉRISTIQUES TEMPORELLES DE LA PRODUCTION ET DE LA PERCEPTION D'ÉMOTION DANS LA PAROLE

Sylvie J. L. MOZZICONACCI[†] et Dik J. HERMES

*IPO (Center for user-system interaction), Eindhoven, Pays-Bas,
†Aussi : Institut de Phonétique d'Amsterdam, Amsterdam, Pays Bas,
et Laboratoire de Phonétique de l'Université de Leiden, Leiden, Pays-Bas
Mel : Mozziconacci@hotmail.com*

Résumé

La présente étude traite des aspects temporels de la parole empreints de l'état émotionnel du locuteur. Tout d'abord, une étude de production a été conduite au niveau global de la phrase entière, portant sur le "débit global". Pour chaque émotion, la valeur moyenne du débit global a été déterminée. Une étude de perception a aussi été effectuée au niveau global. Les valeurs trouvées optimales à l'issue de l'expérience de perception sont comparées aux valeurs moyennes obtenues lors de l'étude de production. Il se peut cependant que certaines informations à un niveau plus détaillé soient spécifiques à certaines émotions et qu'elles permettent de les distinguer. Une étude portant sur le "débit local" a donc été menée. Afin de distinguer les effets temporels dus à l'expression d'une émotion de ceux induits par la variation de débit global de parole, l'étude a porté, dans un premier temps, sur les durées relatives des segments accentués et non-accentués en parole dénuée d'émotion produite à débit variable, et dans un deuxième temps, sur ces durées relatives en parole émotionnelle. La pertinence des différences observées entre parole dénuée d'émotion et parole émotionnelle a été testée lors d'une expérience de perception. Ces variations locales se révèlent être pertinentes pour certaines émotions, alors qu'une manipulation globale de la durée totale est considérée suffisante pour certaines autres émotions.

Abstract

The present study investigates temporal characteristics of speech conveying emotion and attitude. First, a production study was conducted at the global level of the whole utterance, determining the production values of "global speech rate" for each emotion and attitude involved in the study. A perception study was also carried out at this global level, seeking perceptually optimal "global speech rate" values for conveying each emotion and attitude. Perception and production values were compared. Second, as more local information inside utterances might be specific to particular emotions or attitudes, the study went a step further with the analysis of "local speech rate". In order to distinguish the local temporal effect of variation in global speech rate from the effect of conveying an emotion, the study considered accented and non-accented speech segments, in neutral speech at variable speech rate and in emotional speech, separately. The perceptual relevance of the differences observed between neutral and emotional speech was then tested in a perception experiment. Local variation appeared to be relevant for generating some emotions in speech while a linear manipulation appeared sufficient for other emotions. Summarizing, both global and local variations appeared to be relevant for conveying emotions and attitudes in speech.

1. Introduction

La parole, tout en véhiculant le contenu littéral du message parlé, l'accompagne d'informations concernant, entre autres, l'identité du locuteur, son genre, son origine régionale, son milieu social, son état émotionnel et son attitude envers l'interlocuteur, le sujet de conversation et la situation. La prosodie remplit de multiples fonctions communicatives : celle de structurer le message en signalant par exemple les parties importantes du discours, celle d'indiquer la structure de la communication orale, et celle de caractériser le locuteur. Les variations prosodiques, comprenant des variations mélodiques, d'intensité, de qualité de voix, de débit et de rythme de parole contribuent entre autres à l'expression et à la perception de l'émotion dans la parole, ce qui constitue le thème de la présente étude.

Le fait qu'on ne dispose actuellement ni de définition de l'émotion communément acceptée, ni de taxonomie correspondante, pose une difficulté méthodologique. Plutchik (1980) a inventorié différentes définitions de l'émotion proposées par divers auteurs. Le terme *émotion* prend différents sens dans la littérature, et s'étend parfois aux notions d'attitude et d'intention. Bien

que l'émotion puisse être considérée comme un phénomène complexe ayant des aspects biologiques, sociaux et personnels, et l'attitude plutôt comme une disposition du locuteur, nous utilisons dans notre terminologie le terme *émotion* de manière large, l'appliquant, par extension, aussi aux attitudes.

Divers travaux se sont intéressés à la fonction expressive de la prosodie (par exemple Uldall, 1964 ; Abercrombie, 1968 ; Léon, 1976 ; Fónagy, 1983 ; Scherer et *al.*, 1984). Quelques auteurs (Frick, 1985 ; Scherer, 1986 ; Murray et Arnott, 1993) ont passé en revue les études concernant l'expression de l'émotion dans la parole. Divers travaux portent donc sur les variations prosodiques véhiculant l'émotion, mais le nombre d'études quantitatives consacrées à l'aspect temporel de la parole exprimant de l'émotion (par exemple Williams et Stevens, 1972 ; van Bezooijen, 1984 ; Cahn, 1990 ; Kitahara et Tohkura, 1992) est limité, et les efforts se concentrent généralement sur des différences de débit global d'une émotion à l'autre. Le but de la présente étude est de cerner la contribution des variations temporelles à l'expression et la perception de l'émotion dans la parole en néerlandais. Les caractéristiques temporelles inter-émotions et inter-

locuteurs sont décrites aussi bien au niveau global de la phrase entière qu'au niveau local de segments accentués et non-accentués dans des énoncés exprimant une émotion. Au niveau global, les variations en durée totale des phrases sont exprimées en termes de *débit global*, ce qui signifie que seules les variations linéaires consistant en un allongement ou une compression de la durée totale sont prises en compte. Au niveau local, les variations en durée relative des segments de parole accentués et non-accentués sont considérées et exprimées en termes de *proportion accentuée*, ce qui équivaut à une certaine expression de la notion de *débit local*.

De plus, l'approche adoptée dans la présente étude combine études de production et de perception. En effet, il est souhaitable de vérifier la pertinence perceptive des résultats obtenus lors d'une étude de production afin de ne modéliser que les variations temporelles qui sont pertinentes dans le cadre de la communication parlée.

L'intérêt d'une telle étude réside principalement en l'acquisition de connaissances concernant la prosodie, et en particulier la façon dont la prosodie contribue à la réalisation de l'expressivité dans la parole. En outre, les résultats obtenus pourront être bénéfiques dans le domaine de la synthèse de parole afin d'y enrichir et d'y varier la prosodie, ce qui devrait rendre la parole synthétique plus naturelle.

2. Débit au niveau global

2.1. Analyse de parole naturelle

Trois locuteurs néerlandais (deux hommes, H1 et H2, et une femme, F1) ont été enregistrés lors de l'expression des sept émotions : neutralité (en tant que catégorie de référence), joie, ennui, colère, tristesse, peur et indignation. Afin de susciter ces émotions, les locuteurs ont d'abord dit des textes sémantiquement porteurs de l'émotion en question et étant susceptibles d'évoquer des situations émotionnelles. Ils ont ensuite dit les phrases alors qu'ils se sentaient de l'humeur correspondante. Par émotion, les locuteurs ont produit chacun trois énoncés de chacune des cinq phrases suivantes :

(1) *Zijn vriendin kwam met het vliegtuig*

[Son amie venait en avion]

(2) *Jan is naar de kapper geweest*

[Jean est allé chez le coiffeur]

(3) *Het is bijna negen uur*

[Il est presque neuf heures]

(4) *Ze hebben een nieuwe auto gekocht*

[Ils ont acheté une nouvelle voiture]

(5) *De lamp staat op het bureau*

[La lampe est sur le bureau]

Le contenu de ces phrases est considéré sémantiquement neutre. Le matériel utilisé consiste en 315 énoncés (3 locuteurs × 7 émotions × 3 occurrences × 5 phrases).

La durée totale de chaque énoncé a été mesurée. Les cinq phrases n'induisant pas la réalisation de pauses, la notion de *débit global* peut être ici considérée de la façon la plus simple, comme inversement proportionnelle à la durée totale de l'énoncé. Cette notion de débit global est le rapport entre la durée moyenne de la phrase exprimant une émotion et la durée moyenne de la même phrase en parole neutre par le même locuteur. Ainsi, par exemple, la valeur de débit '0.80' correspond à une réduction de débit, c'est-à-dire un allongement de 20% de la durée de la parole émotionnelle par rapport à la parole neutre du même locuteur. Les résultats sont présentés dans le Tableau 1, en fonction du locuteur. Le débit global moyenné sur les 3 locuteurs y figure aussi, ainsi que les écarts-types, qui sont spécifiés entre parenthèses.

Tableau 1 — Débit global par locuteur et débit global moyen. Les écarts-types correspondants figurent entre parenthèses.

Émotion	Locuteur H1	Locuteur H2	Locutrice F1	Débit moyen
Neutralité	1.00 (.14)	1.00 (.12)	1.00 (.14)	1.00 (.13)
Joie	0.99 (.12)	1.00 (.16)	1.03 (.10)	1.01 (.13)
Ennui	0.73 (.20)	0.87 (.19)	0.85 (.17)	0.82 (.18)
Colère	1.16 (.14)	0.84 (.41)	0.82 (.19)	0.94 (.25)
Tristesse	0.93 (.15)	0.87 (.32)	0.96 (.18)	0.92 (.22)
Peur	1.15 (.13)	1.25 (.15)	0.93 (.18)	1.11 (.15)
Indignation	0.86 (.25)	1.00 (.18)	0.68 (.19)	0.85 (.20)

Les résultats obtenus montrent que les 3 locuteurs sont souvent en accord sur le type de variation du débit global pour l'expression des émotions. Ils sont unanimes pour exprimer la joie en utilisant un débit très proche de celui qu'ils adoptent en parole neutre, et pour ralentir leur parole en exprimant la tristesse et l'ennui. Par contre, lors de l'expression de la colère, alors que H1 et F1 ralentissent leur débit, H1 l'accélère. Pour la peur et l'indignation, c'est F1 qui diffère des autres locuteurs en parlant relativement plus lentement qu'eux. Ces quelques différences témoignent de diverses stratégies d'expression pour une même émotion. Malgré ces variations, on observe aussi une certaine systématique quant au type de variation temporelle produit au niveau global lors de l'expression d'émotions spécifiques.

2.2. Expérience de perception

Si l'approche linéaire, dans laquelle le débit global est inversement proportionnel à la durée totale de l'énoncé, n'est pas trop grossière pour l'expression de l'émotion dans la parole, des valeurs de débit global permettant de véhiculer chaque émotion de façon optimale doivent pouvoir être déterminées. C'est ce que vise à faire la présente étude. Si ces valeurs correspondent à celles de l'étude de production, la pertinence communicative de la variation temporelle sera établie pour la parole émotionnelle.

2.2.1. Procédure

Afin d'obtenir des échantillons de parole adéquats pour la présente expérience, une sélection a été effectuée sur la base des performances d'identification correcte de l'émotion qui ont été obtenues lors d'une expérience préliminaire de perception effectuée avec le matériel sonore décrit précédemment. Pour chaque émotion, et pour les deux phrases pour lesquelles les scores d'identification les plus élevés ont été obtenus, l'énoncé ayant suscité les meilleures performances a été retenu. Quatorze énoncés ont ainsi été sélectionnés. Il s'agit d'occurrences des deux phrases suivantes :

(a) *Zijn vriendin kwam met het vliegtuig*

[Son amie venait en avion]

(b) *Zij hebben een nieuwe auto gekocht*

[Ils ont acheté une nouvelle voiture]

Ces phrases ont été enregistrées par le locuteur masculin H1 dans l'expression des sept émotions. Pour chacune des deux phrases, l'énoncé exprimant la neutralité a servi de support aux manipulations.

Afin de transplanter, selon une correspondance optimale, les courbes de *F0* des énoncés émotionnels sur les énoncés-supports de la phrase correspondante, un algorithme de Dynamic Time Warping (DTW) a d'abord été appliqué. Par émotion, la durée des segments de parole neutre a ainsi été manipulée de façon à correspondre à celle des énoncés émotionnels de la même phrase. Ces courbes de *F0* alignées ont ensuite été transplantées des énoncés émotionnels sur les énoncés neutres, ceci selon l'algorithme TD-PSOLA (Verhelst et Borger, 1991). Ensuite, la parole ainsi générée a été manipulée linéairement de façon à lui donner, par compression ou expansion globale, la même durée totale que celle des énoncés émotionnels, ce qui a aussi été effectué grâce à la technique PSOLA. La parole ainsi obtenue a donc la même courbe mélodique que les énoncés émotionnels, mais la qualité de voix, l'énergie et les autres micro-caractéristiques de durée sont celles de l'énoncé neutre. Finalement, une manipulation linéaire de 70, 80, 90, 100, 110, 120, et 130 pour cent de la durée de la parole ainsi obtenue a fourni, pour chaque émotion et chaque phrase, les sept stimuli utilisés dans l'expérience de perception.

Les stimuli ont été présentés en deux blocs ; un par phrase. L'ordre des phrases a été contrebalancé entre les sujets. L'ordre de présentation des stimuli a été varié de façon aléatoire différente pour chaque sujet. Dix sujets ont participé au test. Les sujets pouvaient écouter les sept variantes aussi fréquemment qu'ils le souhaitent, en utilisant un casque pour l'écoute et devaient indiquer par ordre de préférence les trois variantes leur semblant exprimer le mieux une émotion donnée.

2.2.2. Résultats

L'ordre de préférence des sujets a ensuite été transformé en un score, 3 points étant attribués à la variante trouvée la meilleure, 2 points pour le second choix, et 1 point pour la troisième sélection de chaque sujet. Le score moyen a été calculé pour chaque variante, séparément pour les deux phrases.

La variante ayant reçu le score le plus élevé est considérée optimale pour l'expression de l'émotion

donnée. Le débit global optimal correspondant est reporté dans le Tableau 2. Par exemple, un débit de parole accru par rapport à celui utilisé en parole neutre a été considéré adéquat pour l'expression de la colère et de la joie, alors qu'une réduction de débit semble approprié à l'expression de l'ennui et de la tristesse.

Le débit global trouvé optimal en perception diffère de moins de 10% d'une phrase à l'autre pour toutes les émotions, excepté l'indignation.

Tableau 2 — Valeurs trouvées perceptivement optimales pour le débit global par rapport à la neutralité et débit global moyen. Un score de 1 signifie en moyenne une troisième place dans les préférences des sujets, alors que le score moyenné maximal, correspondant au choix préférentiel de tous les sujets, serait égal à 3.

Émotion	Débit global et score par phrase				Débit global moyen
	'Zij hebben een nieuwe auto' gekocht'		'Zijn vriendin kwam met het vliegtuig'		
	Débit Global	Score	Débit Global	Score	
Neutralité	1.00	2.00	1.00	2.40	1.00
Joie	1.18	2.30	1.25	2.00	1.20
Ennui	0.65	1.70	0.69	1.90	0.67
Colère	1.28	1.40	1.25	1.90	1.27
Tristesse	0.79	1.80	0.76	2.20	0.78
Peur	1.09	1.50	1.18	1.50	1.12
Indignation	0.78	1.80	0.94	1.40	0.85

2.2.3. Discussion

Une comparaison des résultats de la présente expérience de perception et de ceux de la précédente étude de production (voir les colonnes de droite des tableaux 1 et 2) met en évidence des ressemblances remarquables. Pour certaines émotions, le débit global moyen observé en production et le débit global optimal en perception sont très similaires. Il faut pourtant noter qu'à plusieurs reprises, les valeurs trouvées optimales à l'issue de l'expérience de perception excèdent les valeurs trouvées dans l'étude de production ; là où les locuteurs réalisent une réduction de débit, la valeur trouvée optimale en perception suggère une réduction plus importante. Ceci est probablement dû au fait que lors de l'expérience de perception, seul le débit global était soumis à variation. En effet, pour s'exprimer en parole naturelle, on fait appel à divers paramètres et leurs effets combinés. Il semble raisonnable que les sujets, ne pouvant se fier qu'au débit pour identifier l'émotion lors du test de perception, aient préféré des valeurs plus importantes qu'en parole naturelle.

De plus, les résultats présents correspondent assez bien à ceux d'études antérieures effectuées dans le même domaine. Pour faciliter la comparaison, les valeurs obtenues dans quelques-unes de ces études ont été converties en valeurs de débit global par rapport à la condition de neutralité. Les présents résultats sont proches de ceux de van Bezooijen (1984). Les valeurs

qu'elle rapporte pour la joie, la colère et la peur sont respectivement 1.20, 1.16 et 0.99, et sont fort comparables à celles obtenues ici grâce au test de perception. La seule différence est que le débit global de 1.25 qu'elle trouve pour la tristesse représente un débit de parole plus rapide que celui trouvé dans notre étude. En ce qui concerne les travaux de Carlson, Granström et Nord (1992), la principale divergence porte sur la parole neutre, qui, dans leur matériel sonore, est plus rapide que lors de l'expression des émotions, ce qui rend la comparaison de débit global par rapport à la neutralité difficile. Une simple comparaison entre elles des valeurs de débit attribuées pour les émotions, montre que dans l'expression de la joie et de la colère, des débits de parole assez similaires sont utilisés, respectivement 0.77 et 0.74, alors que la tristesse est exprimée avec un débit plus réduit, en l'occurrence 0.66. Ceci correspond assez bien à nos propres résultats. Les valeurs de débit proposées par Kitahara et Tokhura (1992), 1.00 pour la joie, 1.43 pour la colère, et 0.87 pour la tristesse sont également en assez bonne correspondance avec les nôtres.

Il peut être conclu que les résultats de la présente étude sont comparables, au moins de façon qualitative, avec ceux d'autres études, certaines ayant porté sur d'autres langues. De plus, la convergence des résultats obtenus en production et en perception confirme la pertinence des variations temporelles globales pour la communication, ainsi que l'estimation des valeurs trouvées les plus adéquates pour l'expression de ces émotions.

3. Analyse de débit au niveau local

Sachant que les variations en débit moyen de parole contribuent à l'expression de l'émotion dans la parole, il semble souhaitable de considérer aussi la façon dont les variations temporelles sont réalisées à l'intérieur des phrases. Il se peut en effet que l'expression de l'émotion exprimée détermine l'organisation temporelle au niveau local. En fonction de l'émotion exprimée, diverses options se présentent pour le choix des unités à considérer. Par exemple, l'allongement en position finale, la proportion de segments de silence dans la phrase, la façon dont les unités phonétiques sont affectées (par des réductions, des assimilations, etc.), et leur durée étaient des candidats à l'étude. Néanmoins, la solution la plus simple, permettant d'étudier les variations temporelles à l'intérieur des phrases, est de considérer la durée relative des segments accentués et non-accentués. Un segment de parole accentué est composé d'une syllabe porteuse de *stress* lexical et sur laquelle un accent mélodique est réalisé. Un segment non-accentué est composé d'une syllabe ou de plusieurs syllabes consécutives sur lesquelles aucun accent n'est réalisé. Pour aller au plus simple et exclure aussi l'effet de l'allongement final, la dernière syllabe des phrases est simplement exclue des analyses.

Afin de distinguer l'effet propre à l'expression d'émotion de celui propre à un simple changement de débit de parole, il est nécessaire de pouvoir comparer la durée de segments de parole émotionnelle à celle de segments de parole dénuée d'émotion produite à des débits variables. Une référence fiable n'a pas pu être trouvée dans la littérature. En effet, pour une réduction de débit, il a été reporté que le temps de parole supplémentaire est réparti de façon égale sur les syllabes accentuées et les syllabes non-accentuées

(Peterson et Lehiste, 1960 ; Kozhevnikov et Chistovich, 1965 ; Miller, 1981). Pour une augmentation de débit, alors que Lehiste (1970) spécifie que la réalisation temporelle dépend du type de langue et peut se réaliser soit par une réduction des syllabes non-accentuées, soit par une réduction égale des syllabes accentuées et non-accentuées, d'autres auteurs reportent qu'il y aurait une tendance à réduire les syllabes non-accentuées relativement plus que les syllabes accentuées (Peterson et Lehiste, 1960 ; den Os, 1988). De plus, n'ayant pas connaissance d'étude concernant la distribution temporelle des segments de parole pour une augmentation progressive du débit, une analyse de parole dénuée d'émotion a été effectuée afin d'obtenir une référence pour le néerlandais. Le but est de déterminer si la proportion de segments de parole accentués et de segments non-accentués est similaire en parole émotionnelle et en parole dénuée d'émotion. Si c'est le cas, un modèle linéaire sera suffisant pour décrire les phénomènes temporels pertinents pour l'expression de l'émotion. Sinon, les règles qui régissent ces variations locales devront être déterminées, et la pertinence de variations locales sera testée pour caractériser les indices perceptifs de l'émotion.

3.1. Matériel

3.1.1. Parole émotionnelle

Parmi les phrases utilisées pour l'analyse globale précédente, deux n'ont pas été considérées adéquates pour la présente analyse. L'une suscite en effet la production d'un accent sur la dernière syllabe de la phrase, de sorte que l'effet propre à l'allongement final interfère avec l'effet temporel propre à l'accentuation. L'autre comporte davantage de syllabes avec *stress* que les autres phrases, c'est-à-dire davantage de syllabes potentiellement porteuses d'accent. Du fait que les locuteurs avaient le choix, ils n'ont pas systématiquement réalisé un accent unique sur la même syllabe de cette phrase. Chacune des trois autres phrases contient deux syllabes avec *stress*. Ci-dessous, les syllabes avec *stress* sont soulignées, les segments accentués sont séparés des non-accentués par des traits verticaux, et les syllabes finales, exclues des analyses figurent en italique.

Les phrases sont les suivantes :

(1) Zijn vrien | din | kwam met het | vlieg | *tuig*

[Zijn vriendin kwam met het vliegtuig]

[Son amie venait en avion]

(2) Jan | is naar de | ka | pper ge | *weest*

[Jan is naar de kapper geweest]

[Jean est allé chez le coiffeur]

(3) Het is | bij | na | ne | gen | *uur*

[Het is bij na negen uur]

[Il est presque neuf heures]

Au total, 182 des 189 énoncés exprimant l'une des sept émotions ont été analysés. Il faut noter que la neutralité étant l'une des sept catégories, 27 de ces énoncés émotionnels expriment la catégorie neutralité. Sept énoncés n'ont pu être considérés, soit que le locuteur ait

bafouillé, soit qu'un allongement ait été produit à une frontière prosodique, soit que deux accents n'aient pas été réalisés dans l'énoncé. Comme ces sept énoncés étaient des réalisations des trois différents locuteurs, exprimant cinq différentes émotions, on peut admettre que le fait de supprimer ces énoncés ne modifie pas sensiblement les résultats.

3.1.2. Parole dénuée d'émotion

Les trois mêmes phrases ont fait l'objet d'un nouvel enregistrement du locuteur masculin H1, produisant, cette fois, de la parole dénuée d'émotion à des débits augmentant progressivement. L'étendue de ces variations en débit couvre celle des valeurs observées en parole émotionnelle. Au total, 171 énoncés ont été enregistrés, c'est-à-dire 57 occurrences par phrase.

3.2. Procédure

Dans les énoncés chargés d'émotion et les énoncés dénués d'émotions produits à débit variable, la durée des segments accentués et non-accentués a été mesurée, du début du segment au début du segment suivant. Par type d'énoncé et par phrase, la durée totale des segments accentués et celle des segments non-accentués (à l'exclusion de la syllabe finale) ont été calculées. La proportion de la durée des segments accentués a été calculée par rapport à la durée totale des segments de cette phrase. Cette proportion sera plus brièvement nommée *proportion accentuée*. Sur la base des valeurs de proportion accentuée obtenues pour les énoncés en parole dénuée d'émotion, la ligne de régression optimale a été calculée pour chaque phrase, afin de constituer la référence de parole dénuée d'émotion. Les données obtenues dans la parole exprimant les émotions sont considérées relativement aux données obtenues dans la parole dénuée d'émotion.

3.3 Résultats et discussion

Les fonctions ci-dessous décrivent les lignes de régression optimales pour les données obtenues dans les énoncés de parole dénuée d'émotion produite à débit variable.

Pour la Phrase 2 (*Jan is naar de kapper geweest*) et la Phrase 3 (*Het is bijna negen uur*), le fait de considérer deux lignes de régression au lieu d'une seule augmente clairement la variation expliquée. La valeur de la racine carrée de la distance moyenne entre ligne de régression et représentation des énoncés diminue de 0.0243, pour une ligne de régression unique, à 0.0227, pour deux lignes de régression en ce qui concerne la Phrase 2 ; et de 0.0229 à 0.0173 pour la Phrase 3 ($F_s = 1.74$, $F_{.05 [56,56]} = 1.56$, $p < .05$). Pour la Phrase 1 (*Zijn vriendin kwam met het vliegtuig*), la différence minime de 0.0174, pour une seule ligne, à 0.0171 pour deux lignes ne justifie pas l'emploi de deux lignes de régression.

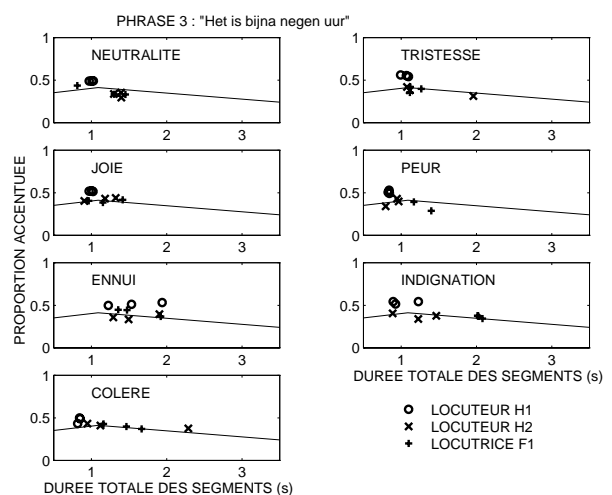
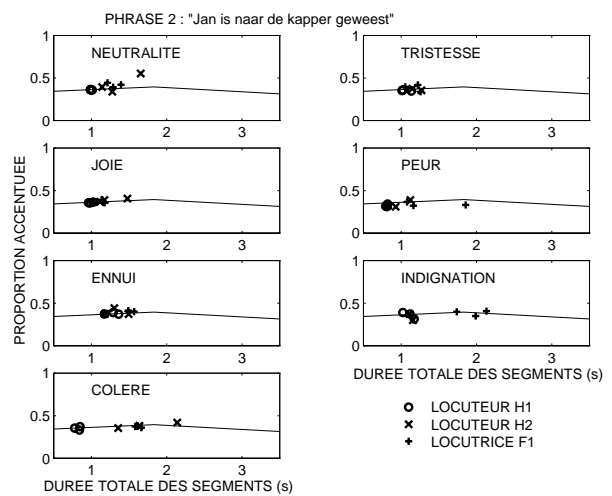
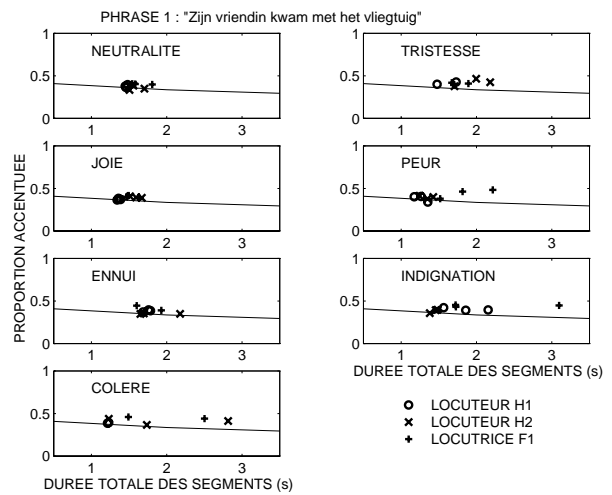


Figure 1 — Proportions de durée des segments de parole accentués, présentées par phrase et par locuteur

Les fonctions suivantes, où x est la durée totale des segments en secondes, et y la proportion accentuée, décrivent les lignes de régression et sont représentées dans la Figure 1.

Phrase 1 : Pour tout x , $y = 0.4206 - (0.0388 \times x)$

Phrase 2 : Pour tout $x < 1.83$, $y = 0.3246 + (0.0387 \times x)$
 autrement, $y = 0.4845 - (0.0487 \times x)$

Phrase 3 : Pour tout $x < 1.09$, $y = 0.3012 + (0.1026 \times x)$
 autrement, $y = 0.4906 - (0.0712 \times x)$

En parole dénuée d'émotion, pour les phrases 2 et 3 et un débit global relativement rapide, plus un ralentissement de la parole est important, plus il provoque un allongement des segments qui touche les segments de parole accentués relativement davantage que les segments non-accentués. Passé un certain point, plus le ralentissement est important moins il privilégie l'allongement des segments accentués, ce qui est aussi le cas pour la phrase 1.

Les résultats concernant les proportions accentuées dans la parole exprimant les émotions sont aussi présentés dans la Figure 1, séparément pour chaque phrase. Les points situés au-dessus des lignes de régression indiquent un allongement des syllabes accentuées plus important qu'en parole dénuée d'émotion. Le cas se présente par exemple pour l'expression de la colère, la tristesse, la peur et l'indignation sur la Phrase 1. Toutefois, la validité de la présente approche est quelque peu restreinte du fait que la référence de parole dénuée d'émotion produite à débit variable ne prend en compte qu'un seul locuteur, H1. De plus, sur la Phrase 3, ce locuteur a produit différentes proportions accentuées dans la parole dénuée d'émotion, et celle exprimant la neutralité dans l'enregistrement de parole émotionnelle. Cette différence semble aussi se répercuter sur l'expression des émotions sur cette même phrase.

Bien que l'analyse mette en évidence des variations en proportions accentuées d'une émotion à l'autre, et que, pour un débit global réduit, on perçoive une certaine tendance à allonger, en parole émotionnelle, les segments accentués relativement plus que les segments non-accentués, les résultats ne permettent pas la formulation d'hypothèses concernant la réalisation exacte des variations locales temporelles pour l'expression d'émotions spécifiques. Cela n'exclut pas non plus la possibilité que la répartition temporelle des segments de parole accentués et non-accentués ne joue aucun rôle dans la communication de l'émotion dans la parole. Dans la suite de cette étude, on a recours à la perception pour tenter d'obtenir davantage d'informations sur la pertinence communicative de la répartition des segments de parole pour l'expression des sept émotions.

4. Identification de l'émotion en fonction du débit au niveau local

L'étude de production décrite précédemment met en évidence des variations locales sans pour autant être concluante en ce qui concerne l'influence de l'expression d'émotion sur la répartition des durées de segments de parole accentués et non-accentués. Par conséquent, la présente étude se propose de tester si les variations locales sont pertinentes pour la perception de l'émotion. Si c'est le cas, les valeurs de proportion accentuée permettant d'exprimer chaque émotion de façon optimale seront déterminées.

4.1. Matériel

Trois énoncés neutres, utilisés dans les analyses précédentes, ont été repris et manipulés par analyse-resynthèse. Il s'agit d'énoncés du locuteur H1 disant chacune des trois phrases utilisées précédemment. Deux séries de stimuli ont été générées, par manipulation de ces énoncés de phrases entières. Chacune des deux séries est constituée des six mêmes conditions.

Dans la Série 1, la durée totale des stimuli est constante. Seul le débit *local* varie en fonction des six conditions. Dans la Série 2, la durée totale des stimuli varie suivant les valeurs de débit *global* trouvées optimales pour chacune des sept émotions dans l'étude précédente (voir la colonne de droite du Tableau 2). De plus, le débit *local* varie en fonction des six conditions.

La Condition 1 ne présente aucune manipulation temporelle au niveau local, laissant la répartition temporelle des segments de parole accentués et non-accentués telle qu'elle était dans les énoncés neutres. La proportion accentuée est variée dans les 5 autres conditions.

Pour la Condition 2, la proportion accentuée dépend entièrement du débit global, et est manipulée suivant les fonctions décrivant les lignes de régression représentées Figure 1 et issues de l'analyse de parole dénuée d'émotion produite à débit variable.

Dans les Conditions 3, 4 et 5, la proportion accentuée est manipulée de façon à être respectivement réduite de 20%, augmentée de 20% et de 40% par rapport à la proportion accentuée "de référence" utilisée pour la Condition 2.

Tous les stimuli des Conditions 1 à 5 ont été générés avec la même configuration intonative ('intonation pattern'), ceci afin de contrôler la variation en type de contour mélodique. Suivant l'approche de l'intonation de l'IPO (t Hart et al., 1990), les configurations intonatives sont les catégories mélodiques abstraites sous-jacentes aux courbes intonatives considérées dans le détail de leur réalisation. Selon la terminologie correspondante, un mouvement intonatif est considéré *audible* s'il est d'une part *perceptible*, c'est-à-dire décelable lors d'une écoute analytique, et d'autre part pertinent pour la communication lors d'une écoute globale telle qu'elle est effectuée lors de dialogues dans la vie quotidienne (t Hart et Collier, 1975). La configuration intonative utilisée dans les 5 premières conditions correspond à la réalisation d'une montée et d'une descente mélodique sur chacune des deux syllabes accentuées. En termes de la grammaire intonative de l'IPO, qui prend en compte les mouvements mélodiques et dans laquelle les montées mélodiques peuvent être transcrites par des chiffres et les descentes mélodiques par des lettres, il s'agit de la configuration '1&A 1&A'. Le '1' représente une montée mélodique réalisée tôt et rendant la syllabe prédominante, le 'A' une descente mélodique tardive, et le '&' indique que les deux mouvements sont réalisés sur une même syllabe. La configuration intonative '1&A 1&A' s'est révélée pouvoir être considérée acceptable dans l'expression de toutes ces émotions (Mozziconacci, 1998). Il faut cependant noter que l'allongement de la voyelle, et tout particulièrement un allongement prononcé tel que celui effectué pour la Condition 5, a pour effet de rendre la descente mélodique 'A' plus

clairement audible. Ceci pourrait influencer l'identification d'émotion. Afin de distinguer l'effet d'allongement de la syllabe de celui de l'audibilité du mouvement mélodique, une sixième condition a donc été ajoutée.

Dans la Condition 6, non seulement la proportion accentuée a été accrue de 40%, comme dans la Condition 5, mais la configuration intonative standard '1&A 1&A' a été remplacée par '1B 1B'. La descente mélodique 'B' ne rend pas la syllabe proéminente, contrairement au mouvement 'A'. Le 'B' a été synthétisé de façon à commencer à la fin de la voyelle de la syllabe accentuée, ceci afin que le mouvement mélodique soit à peine audible, sinon inaudible.

Dans toutes les conditions, les valeurs de *pitch level* et de *pitch range* trouvées optimales dans une étude préalable (Mozziconacci, 1998) ont été utilisées. De plus, la syllabe finale des phrases n'a pas été affectée par les manipulations, pour éviter une éventuelle interférence avec l'allongement final. Toutes les manipulations ont été basées sur la technique PSOLA (Moulines et Laroche, 1995). Au total, 252 stimuli (2 séries × 6 conditions × 3 phrases × 7 combinaisons de *pitch level* et *pitch range*) ont été produits. Les manipulations de durée correspondant aux 6 conditions sont récapitulées dans le Tableau 3.

4.2. Procédure

Vingt-quatre sujets ont participé à ce test de perception dans lequel les stimuli leur étaient présentés en 3 blocs (un par phrase) dans un ordre aléatoire différent pour chaque sujet. Après avoir écouté le stimulus une fois, ils devaient choisir, parmi les sept catégories d'émotions qui leur étaient proposées, celle qui selon eux avait été exprimée.

Tableau 3 — Récapitulation des conditions qui correspondent aux manipulations au niveau local. "Proportion accentuée" signifie que cette proportion accentuée dépend entièrement du débit global.

	Manipulation temporelle au niveau local	Configuration intonative
Cond. 1	Aucune	1&A 1&A
Cond. 2	Proportion accentuée	1&A 1&A
Cond. 3	Proportion accentuée -20%	1&A 1&A
Cond. 4	Proportion accentuée +20%	1&A 1&A
Cond. 5	Proportion accentuée +40%	1&A 1&A
Cond. 6	Proportion accentuée +40%	1B 1B

4.3. Résultats

Les résultats des tests d'identification ont subi une analyse log-linéaire (Fienberg, 1980). Cette analyse présente l'avantage de prendre en compte les informations concernant les confusions dans les réponses des sujets, au lieu de ne prendre en considération que leurs réponses correctes. Une série de prédictions est calculée, concernant les réponses des sujets en fonction de la présence ou de l'absence d'effet significatif des variables et de leurs interactions. Ces

prédictions sont exprimées sous forme de modèles log-linéaires qui peuvent être comparés aux réponses obtenues des sujets.

Dans notre étude, le modèle log-linéaire décrivant le mieux les données est celui où il existe des effets significatifs de PITCH (les combinaisons de *pitch level* et *pitch range*), COND (les conditions), RESP (les réponses des sujets), et des interactions significatives entre PITCH et RESP, et entre COND et RESP.

Une analyse de *clusters* a permis de former deux *clusters*, chacun de trois conditions. Le premier *cluster* réunit les trois premières conditions, correspondant à de faibles proportions accentuées. Les trois dernières conditions forment le *cluster* des proportions accentuées les plus élevées. Le fait que les Conditions 5 et 6 fassent partie du même *cluster* indique que l'effet qui influe sur les réponses des sujets est bien lié à la longueur des syllabes et non à l'audibilité du mouvement mélodique. Les résultats sont présentés dans le Tableau 4 pour les deux *clusters* ainsi formés. Les flèches indiquent que le nombre de réponses des sujets dans la catégorie correspondante est, de façon significative, plus élevé (↑) ou moins élevé (↓) que selon les prédictions d'un modèle log-linéaire représentant l'absence d'effet des conditions sur ces réponses. Des déviations significatives de ce modèle ne sont obtenues que pour les catégories neutralité et indignation, mais dans ces deux cas, ces variations temporelles s'avèrent très importantes pour la perception. La proportion accentuée doit être limitée en parole neutre. En effet, une augmentation de la proportion accentuée se fait au détriment de la perception de neutralité. En l'occurrence, une proportion accentuée qui est importante suggère la perception de l'indignation.

Tableau 4 — Nombre de réponses par *cluster* de conditions.

Conditions	Réponses des sujets						
	Neutr.	Joie	Ennui	Colère	Trist.	Peur	Indign.
Série 1							
Cond. 1 à 3	566↑	290	63	98	187	146	162↓
Cond. 4 à 6	387↓	259	102	92	206	159	307↑
Série 2							
Cond. 1 à 3	280↑	236	328	179	184	136	169↓
Cond. 4 à 6	169↓	223	338	174	160	159	289↑

Afin de mieux distinguer l'effet du débit global de celui du débit local, une autre analyse des résultats a été effectuée. Pour cette analyse, les performances d'identification ont été considérées suivant le *type de stimulus*. Les stimuli utilisés dans les deux séries de six conditions ont été répartis en quatre types de stimuli. Les stimuli de la Série 1, générés avec un débit *global* constant, constituent les stimuli de Types 1 et 2, alors que les stimuli de la Série 2, générés avec un débit *global* variant suivant les valeurs optimales, constituent les stimuli de Types 3 et 4. Le débit *local*, constant pour les stimuli de Types 1 et 3, distingue ces stimuli de ceux de Types 2 et 4, se caractérisant par un débit *local*

variant suivant les valeurs optimales. Les pourcentages d'identification correcte pour ces 4 différents types de stimuli figurent dans le Tableau 5. Il n'est pas étonnant que les pourcentages d'identification correcte soient élevés pour la neutralité, puisque les stimuli ont été générés à partir d'énoncés neutres et qu'ils sont porteurs de traits mélodiques et d'une qualité vocale caractérisant la neutralité. Les scores des catégories joie et peur sont peu améliorés par les types de conditions. Par contre, pour la colère et surtout pour l'ennui, la variation en débit global se révèle être d'une grande importance, alors que pour la tristesse et l'indignation, c'est la variation en débit local qui apparaît influencer le plus sur les scores. Quoi qu'il en soit l'effet le plus marquant est celui de la réduction du débit global pour l'expression de l'ennui.

Le pourcentage correct moyen figurant dans la colonne de droite du Tableau 5 fournit de nouvelles informations sur les effets respectifs du débit global et du débit local. En ce qui concerne l'effet du débit global, la comparaison des résultats obtenus pour les stimuli de Types 1 (débit *global* constant) et 3 (débit *global* optimal) met en évidence une amélioration de 17% de la performance d'identification, alors que la comparaison des résultats obtenus pour les stimuli de Types 2 (débit *global* constant) et 4 (débit *global* optimal) met en évidence une amélioration de 16%. Quant à l'effet du débit local, comparer les résultats obtenus pour les stimuli de Types 1 (débit *local* constant) et 2 (débit *local* optimal) nous mène à constater une amélioration de 8%, alors qu'une amélioration de 7% apparaît en comparant les stimuli de Types 3 (débit *local* constant) et 4 (débit *local* optimal). En moyenne, la modélisation du débit global permet d'améliorer l'identification des émotions de 16.5%, et celle du débit local d'améliorer ces performances de 7.5%. De plus, les deux effets semblent bien s'additionner comme le montre la comparaison des Types 1 et 4, ce qui suggère que ces effets sont indépendants.

Tableau 5 — Pourcentages d'identification correcte.

Type 1 : débit global constant, débit local constant
 Type 2 : débit global constant, débit local optimal
 Type 3 : débit global optimal, débit local constant
 Type 4 : débit global optimal, débit local optimal

	Neutr.	Joie	Ennui	Colère	Trist.	Peur	Indign.	Moyenne
Type 1	74	28	13	1	13	29	10	24
Type 2	74	32	22	8	21	35	33	32
Type 3	76	36	85	31	18	25	13	41
Type 4	76	36	89	33	31	32	36	48

5. Discussion

Il a été confirmé que les variations temporelles au niveau global de la phrase entière sont d'une importance primordiale pour la communication de l'émotion dans la parole, ceci tout particulièrement pour certaines émotions comme l'ennui. Comparé à l'effet majeur du débit global ou celui de la mélodie, celui des variations en proportion accentuée est relativement limité, ce qui n'a rien d'étonnant.

Une généralisation de ces résultats ne pourrait être faite que prudemment, à cause des limitations de cette étude qui ne comporte que le nombre limité de trois locuteurs, s'exprimant tous dans la même langue, le néerlandais, et exprimant des émotions qui ne sont pas spontanées, mais ont été suscitées dans une situation de laboratoire. Il sera intéressant de vérifier si les effets temporels dont la pertinence pour l'expression d'émotion a été mise en évidence au niveau local, c'est-à-dire dans le cours des phrases, se retrouveront dans de la parole émotionnelle recueillie en situation spontanée. Une étude multilingue permettrait d'ajuster de tels paramètres acoustiques pour des langues spécifiques, et d'observer la pertinence relative des paramètres d'un type de langue à l'autre.

Bien que l'analyse de production n'ait montré qu'une tendance à varier la structure temporelle lors de l'expression d'émotion, le test de perception a démontré que les variations en proportion accentuée remplissent une fonction communicative, telle que celle de signaler certaines émotions dans la parole. Ceci met bien en évidence l'intérêt d'utiliser de manière complémentaire les études en production et les études en perception. Une telle approche permet de rendre compte de la variabilité observée dans les études de production. Mais elle peut aussi révéler la pertinence communicative d'une variable. Il s'avère que des variables autres que celles étudiées co-varient naturellement dans les énoncés soumis aux analyses. Grâce au contrôle exercé sur ces variations dans les tests de perception, en donnant des valeurs constantes aux variables autres que celles sur lesquelles porte l'étude, l'effet de la variable étudiée peut être déterminé. Ainsi, l'effet du débit local qui n'émergeait pas statistiquement dans l'étude de production s'est révélé significatif lors de l'expérience de perception.

Cette complémentarité des études de production et de perception a été exploitée dans la présente approche, afin de déterminer comment la prosodie véhicule l'émotion. L'approche choisie vise à cerner d'abord des effets de magnitude relativement importante, ce qui peut être fait grâce à l'analyse de la parole émotionnelle d'un nombre limité de locuteurs. Des expériences de perception permettent ensuite de vérifier si ces effets sont bien pertinents pour la communication orale. Cette approche a été utilisée de manière itérative. Dans un premier temps, elle a été employée au niveau global de la phrase entière, pour traiter des effets les plus importants. Dans un second temps, elle a été réutilisée au niveau local, considérant la proportion des segments de parole accentués et non-accentués. L'étape suivante consisterait à poursuivre les recherches à un niveau plus détaillé, ce qui nécessiterait une étude de production à plus grande échelle, analysant la parole émotionnelle d'un large nombre de locuteurs. Une telle étude se concentrerait plutôt sur la variabilité inter-locuteurs, et permettrait mieux de décrire des effets de magnitude moins importante que ceux décrits précédemment. De nouveau, la pertinence de ces effets pour communiquer l'émotion, ne pourra être établie sans expérience de perception.

Remerciements

Nous tenons à remercier Jacqueline Vaissière pour ses précieux commentaires ayant contribué à l'amélioration

de cet article. De plus, les modifications d'une version antérieure ont eu lieu durant la participation du premier auteur au projet CREST *Emotional speech* (Japan Science and Technology).

Références bibliographiques

[**Abercrombie, 1968**] Abercrombie D. (1968). Paralanguage, *British Journal of Disorders of Communication*. 3. 55-59. Reprinted in D. Abercrombie (ed.) (1991), *Fifty years in Phonetics: selected papers*. Edinburgh University press: Edinburgh. 101-108.

[**van Bezooijen, 1984**] van Bezooijen R. A. M. G. (1984). The characteristics and the recognizability of vocal expression of emotion. Foris, Dordrecht, The Netherlands.

[**Cahn, 1990**] Cahn J. E. (1990). Generating expression in synthesized speech. Technical report, MIT Media Lab., Boston.

[**Carlson et al., 1992**] Carlson R., Granström B., Nord L. (1992). Experiments with emotive speech: acted utterances and synthesized replicas. *Proceedings ICSLP 92*. Banff, Alberta, Canada. 1. 671-674.

[**Fienberg, 1980**] Fienberg S. E. (1980). The analysis of cross-classified categorical data. MIT Press: Cambridge, Massachusetts.

[**Fónagy, 1983**] Fónagy I. (1983). La vive voix: Essais de psycho-phonétique. Payot: Paris.

[**Frick, 1985**] Frick R. W. (1985). Communicating emotion: the role of prosodic features. *Psychological Bulletin*. 97. 412-429.

[**'t Hart et Collier, 1975**] 't Hart J., Collier R. (1975). Integrating different levels of intonation analysis. *Journal of Phonetics*. 3. 235-255.

[**'t Hart et al., 1990**] 't Hart J., Collier R., Cohen A. (1990). A perceptual study of intonation. Cambridge University Press: Cambridge.

[**Kitahara et Tohkura, 1992**] Kitahara Y., Tohkura, Y. (1992). "Prosodic control to express emotions for man-machine interaction". *IEICE Transactions on Fundamentals of Electronics, Communication and Computer Sciences*. 75. 155-163.

[**Kozhevnikov et Chistovich, 1965**] Kozhevnikov V. A., Chistovich L. A. (1965). Speech: articulation and perception (Translation). Joint Publications Research Service. Washington D. C.

[**Lehiste, 1970**] Lehiste I. (1970). Suprasegmentals. MIT Press: Cambridge, Massachusetts.

[**Léon, 1976**] Léon P. (1976). De l'analyse psychologique à la catégorisation auditive et acoustique des émotions dans la parole. *Journal de Psychologie*. 3-4.

[**Miller, 1981**] Miller J. L. (1981). Effects of speaking rate on segmental distinctions. In Eimas P. D., Miller J. L. (eds.) *Perspectives on the study of speech*. Lawrence Erlbaum Associates: Hillsdale, New Jersey. 39-74.

[**Moulines et Laroche, 1995**] Moulines E., Laroche J. (1995). Non-parametric techniques for pitch scale and time-scale modification of speech. *Speech Communication*. 16. 175-205.

[**Mozziconacci, 1998**] Mozziconacci S. J. L. (1998). Speech variability and emotion: production and perception. Eindhoven.

[**Murray et Arnott, 1993**] Murray I. R., Arnott J. L. (1993). Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*. 93. 1097-1108.

[**den Os, 1988**] den Os E. (1988). Rhythm and tempo of Dutch and Italian; a contrastive study. Elinkwijk, Utrecht, The Netherlands.

[**Peterson et Lehiste, 1960**] Peterson G. E., Lehiste I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*. 32. 693-703.

[**Plutchik, 1980**] Plutchik, R. (1980). Emotion: a psychoevolutionary synthesis. Harper & Row: New York.

[**Scherer et al., 1984**] 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 (5). 1346-1356.

[**Scherer, 1986**] Scherer K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological Bulletin*. 99. 143-165.

[**Uldall, 1964**] Uldall E. (1964). Dimensions of meaning in intonation. In Abercrombie D., Fry D. B., MacCarthy P. A. D., Scott N. C., Trim J. L. M. (eds.) *In honour of Daniel Jones, Papers contributed on the occasion of his eightieth birthday*. Longman. 271-279. Reprinted in Bolinger D. (ed.) (1972) *Intonation: selected readings*. Penguin Books: Harmondsworth, England. 250-259.

[**Verhelst et Borger, 1991**] Verhelst W., Borger M. (1991). Intra-speaker transplantation of speech characteristics: an application of waveform vocoding techniques and DTW. *Proceedings Eurospeech'91*. Genova, Italy. 3. 1319-1322.

[**Williams et Stevens, 1972**] Williams C. E., Stevens K. N. (1972). Emotions and speech: some acoustical factors. *Journal of the Acoustical Society of America*. 52. 1238-1250.

Les auteurs

Sylvie Mozziconacci a fait des études d'orthophonie à l'université Paris VI (France), puis des études de linguistique informatique à l'université d'Amsterdam (Pays-Bas). Durant la préparation de son doctorat à l'Institut de Recherches sur la Perception (IPO, Eindhoven, Pays-Bas), elle a aussi travaillé au KTH (Stockholm, Suède). Sa thèse porte sur la variabilité prosodique véhiculant l'émotion dans la parole. Elle a ensuite coordonné le projet plurifacultaire sur la prosodie à l'université de Genève (Suisse). Elle continue actuellement ses études post-doctorales à l'université de Leiden (Pays-Bas), toujours sur le thème de l'expression et de la perception de l'émotion et de l'attitude dans la parole. Elle participe actuellement au projet CREST (Sciences et Technologie au Japon), en collaboration avec l'Institut de la Communication Parlée (Grenoble, France).



Dik Hermes a étudié la biophysique à l'université d'Amsterdam (Pays-Bas). Il a travaillé sur le traitement de l'information dans le système auditif des vertébrés. Ces travaux, qui ont mené à un doctorat, ont eu lieu à l'université catholique de Nimègue (Pays-Bas). Il travaille depuis à l'Institut de Recherches sur la Perception (IPO, Eindhoven, Pays-Bas), où il a d'abord conduit des recherches sur la parole, s'intéressant principalement à la prosodie, et plus particulièrement au rythme et à l'intonation. Il consacre actuellement ses recherches au traitement du son et s'intéresse aux aspects multi-sensoriels du traitement de l'information.



