

Identification des émotions en voix naturelle et synthétique : paradigme d’ancrage

Ioulia Grichkovtsova¹, Michel Morel¹ et Anne Lacheret²

¹Laboratoire CRISCO, Université de Caen
Esplanade de la Paix, 14032 Caen Cedex, France

²Laboratoire MODYCO, Institut universitaire de France, Paris
Université Paris X, Nanterre
ioulia.grichkovtsova@unicaen.fr

ABSTRACT

The two main objectives of our study were to identify the perceptive role of intonation and voice quality in the identification of emotions and to determine gating points for each studied emotion. We built a multi-speaker corpus, which was used for the development of a new perception test on the basis of gating paradigm and transplantation paradigm. The results we obtained are discussed in the light of their relevance for the voice synthesis of affective speech.

Keywords: affective prosody, emotions, speech synthesis, gating paradigm.

1. INTRODUCTION

La valeur distinctive de l’intonation et de la qualité vocale n’est toujours pas établie de manière fiable pour l’identification des émotions. L’intérêt de cette problématique n’est pas seulement théorique : une meilleure modélisation des émotions permettra des avancées importantes en TAP, en particulier en synthèse de la parole.

La difficulté d’obtenir des résultats concluants est au moins en partie due aux types de corpus utilisés généralement pour l’analyse de l’encodage des émotions. Il en existe deux types principaux : les phrases avec ou sans contenu lexical.

Une combinaison inventée des syllabes présentes dans la langue [1], ou un énoncé dans une langue étrangère, inconnue des locuteurs et des auditeurs [2], ont pu être utilisés dans les expériences perceptives sur la base de phrases non lexicales. Ces études ont montré une influence significative, mais modeste, de l’intonation sur la perception de certaines émotions, contrairement à la qualité vocale dont l’importance pour l’identification des émotions a été parfaitement démontrée [3].

L’utilisation de phrases dépourvues de sens lexical peut être problématique dans l’étude de l’intonation. En effet, notre hypothèse est que sans la distinction mots pleins/mots vides, sans les frontières syntaxiques et les frontières informationnelles, les points d’ancrage pour les patrons prosodiques qui encodent l’expression des

émotions sont affectés et le rôle de l’intonation est minimisé.

Les études [4, 5] utilisant des phrases dotées d’un sens lexical mettent en lumière l’importance de l’intonation dans l’identification des émotions, même si cette importance est variable. La présence de l’intonation naturelle peut même être cruciale pour certains états affectifs. Néanmoins, cette approche provoque un autre biais : la neutralisation complète du lexique émotionnel n’est pas possible. L’utilisation d’un patron prosodique, même "neutre", préserve toujours la mise en relief des mots importants du lexique et de ce fait facilite l’identification des émotions.

Même si quelques corpus multi-locuteurs sont apparus [1], une autre faiblesse méthodologique est souvent présente dans les expériences sur l’identification des émotions : l’insuffisance du nombre de locuteurs (1 à 3). Or, la parole émotionnelle est caractérisée par une variabilité inter- et intra- individuelle significative [6]. La question se pose de savoir si cet échantillon des locuteurs est représentatif de l’ensemble des stratégies phonostylistiques possibles pour l’encodage des émotions.

2. METHODOLOGIE

2.1. Établissement du corpus

L’analyse des problèmes méthodologiques soulevés en *supra* nous a conduits à créer un corpus multi-locuteur sur la base d’une phrase émotionnellement neutre : *Je vais rentrer à la maison maintenant*. Vingt-deux locuteurs natifs du français (11 hommes et 11 femmes) ont été enregistrés. La taille du corpus représente environ 7 heures d’enregistrement. Six émotions (colère, tristesse, joie, dégoût, peur, chagrin) et une affirmation neutre ont été sélectionnés pour ce travail. S’il était techniquement difficile de tester plusieurs types de phrases neutres au cours de la même expérience, rien n’interdit de reprendre celle-ci avec d’autres phrases syntaxiquement et rythmiquement différentes afin d’affiner nos résultats.

Pour chaque émotion, un texte spécifique devait être oralisé mais à chaque fois avec l’insertion de la même phrase neutre au milieu du texte. Il a été ainsi possible de récolter la même séquence lexicalement neutre mais

dans des contextes émotionnels variés. L'hypothèse sous-jacente était la suivante : l'expression des émotions induites dès les premiers mots du texte, au contenu lexical émotionnellement marqué, se propage dans la production de cette phrase dite neutre ; autrement dit, la phrase en question porte les traces prosodiques de l'émotion globale véhiculée dans le texte, sans que le locuteur en soit conscient.

Chaque texte a été lu trois fois par chaque locuteur pour que celui-ci imagine mieux le contexte de la situation et exprime l'émotion désirée le plus naturellement possible¹.

2.2. Validation du corpus

Le corpus enregistré (154 stimuli présentés dans un ordre aléatoire, 20 minutes par passation) a été validé par un test d'évaluation avec le logiciel PERCEVAL [7]. Dix auditeurs français (moyenne d'âge 30 ans) ont participé au test.

Nous avons fixé le seuil d'acceptation des énoncés identifiées à 50 % (au moins 50 % des auditeurs ont pu identifier l'émotion exprimée). Les résultats du test perceptif sont présentés dans le Tableau 1.

Table 1 : Taux du nombre d'énoncés identifiés par les auditeurs avec le seuil d'acceptation de 50 %. Par exemple, pour la joie, 36 % des énoncés ont été reconnus par au moins 50 % des sujets.

Emotions	Identification réussie
Chagrin	32 %
Colère	73 %
Dégoût	9 %
Joie	36 %
Peur	14 %
Tristesse	68 %

A l'issue de ce test préliminaire, nous avons mis à l'écart le dégoût et la peur, et retenu pour l'analyse les 4 émotions suivantes : chagrin, tristesse, colère et joie.

2.3. Test perceptif

Le corpus multi-locuteur ainsi validé a été utilisé pour le développement d'un nouveau test perceptif qui inclut deux paradigmes méthodologiques : le *paradigme d'ancrage* et le *paradigme de transplantation*.

Paradigme d'ancrage

Le paradigme d'ancrage [7] permet de comprendre quel type et quelle quantité d'information phonétique est nécessaire pour une reconnaissance optimale des émotions. Ce paradigme est fondé sur l'hypothèse selon laquelle un point d'ancrage correspond à une zone de l'énoncé qui déclenche la reconnaissance de l'émotion. La méthodologie est la suivante : le stimulus vocal est

découpé en fragments qui augmentent progressivement par pas de 1 ou 2 syllabes jusqu'à l'énoncé entier² :

je vais | ren | trer | à la | mai | son | main | tenant |

Il est ensuite demandé à chaque auditeur d'identifier l'état affectif à la fin de chaque segment.

Paradigme de transplantation

Le paradigme de transplantation [8] a été utilisé pour évaluer les rôles perceptifs respectifs de la qualité vocale et de l'intonation pour l'identification des 4 émotions (chagrin, tristesse, colère et joie) ainsi que de la production neutre.

Cette méthode repose sur une extraction et un échange du contour prosodique entre deux énoncés avec le même contenu segmental, l'un naturel, l'autre synthétique avec intonation neutre. La transformation prosodique agit sur la fréquence fondamentale, l'intensité et les paramètres temporels, mais elle ne modifie pas, en principe, les caractéristiques de la qualité vocale.

Trois versions de chaque phrase ont été produites selon cette méthode de la transplantation avec le logiciel de synthèse vocale Kali [10] : version 1 « naturelle » (non modifiée), version 2 « qualité vocale » (qualité vocale naturelle de l'acteur et intonation de Kali), version 3 « intonation » (intonation de l'acteur et qualité vocale de Kali).

Au total, 1560 stimuli ont été construits. Compte tenu du nombre de stimuli, le test est basé sur le principe du carré latin³ qui permet de réduire le nombre de stimuli présentés à chaque sujet. Sur ce principe, huit versions du test ont été préparées. Pour sa validation statistique, le test a nécessité 16 sujets (auditeurs français, âge moyen 25 ans, 20 minutes de passation par sujet).

Les stimuli ont été présentés par groupes de segments en commençant par le plus petit, l'ordre des stimuli étant aléatoire dans chaque groupe.

3. RESULTATS

Afin de réduire les incertitudes tout en conservant une précision satisfaisante, un filtre de lissage étroit a été appliqué sur les courbes, chaque valeur V_i étant la moyenne pondérée des valeurs V_{i-1} , V_i et V_{i+1} selon la formule :

$$V_i = 0,25 \cdot V_{i-1} + 0,5 \cdot V_i + 0,25 \cdot V_{i+1}$$

² Nous n'avons pas retenu la méthode temporelle classique (augmentation par rapport à un seuil de durée déterminée) de manière à contourner les variations interlocuteur de débit.

³ Un **carré latin** est un tableau carré de n lignes et n colonnes rempli de n éléments distincts dont chaque ligne et chaque colonne contient un seul exemplaire.

¹ Un exemple de texte est donné en appendice.

Les catégories à identifier étant au nombre de 5, le taux d'identification de 20 % représente le "zéro", c'est-à-dire une valeur purement due au hasard qui servira de référence pour évaluer la significativité des résultats. Les intervalles de confiances à 95 % sont indiqués sur toutes les courbes afin de visualiser les écarts plus ou moins significatifs.

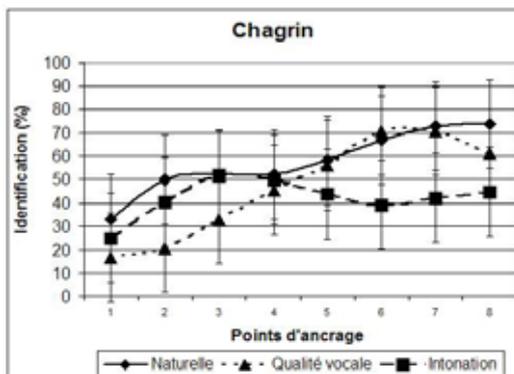


Figure 1 : Chagrin. Version 1 « naturelle », version 2 « qualité vocale », version 3 « intonation ».

Le *chagrin* est caractérisé par une qualité vocale spécifique, Figure 1. Une progression régulière dans l'identification de la version « qualité vocale » est observée jusqu'à un plateau atteint au sixième segment. L'intonation joue un rôle plus modeste, bien que significatif.

Ce premier résultat était prévisible, le chagrin étant une émotion peu contrôlée, qui, en conséquence, modifie d'abord les paramètres physiologiques non codés en langue et influence essentiellement la qualité vocale.

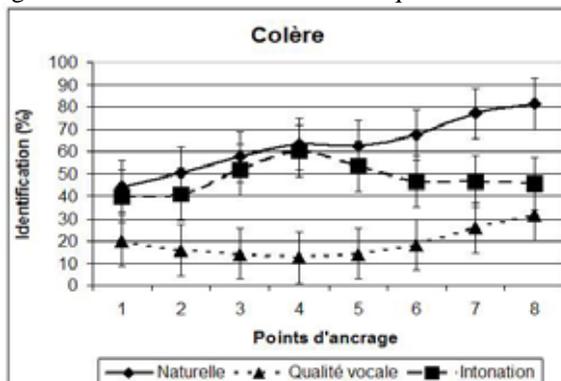


Figure 2 : Colère. Version 1 « naturelle », version 2 « qualité vocale », version 3 « intonation ».

La *colère* met en évidence une bonne identification pour la version naturelle, Figure 2. La qualité vocale joue un rôle plus faible en comparaison avec l'intonation. L'importance de l'intonation fait de la colère un mécanisme socialement et donc linguistiquement contrôlé : relevant plus de l'attitude que de l'émotion.

L'identification de la *joie* naturelle est très supérieure aux versions modifiées, Figure 3. L'intonation est à peine significative, la qualité vocale négligeable. Le mauvais résultat de la qualité vocale résulte de la transplantation prosodique qui altère la voix souriante

dans la version 2 « qualité vocale » et masque la joie. Ce résultat montre les limites de la méthode par transplantation, mais il montre aussi une certaine valeur de l'intonation dans la synthèse d'une émotion telle que la joie.

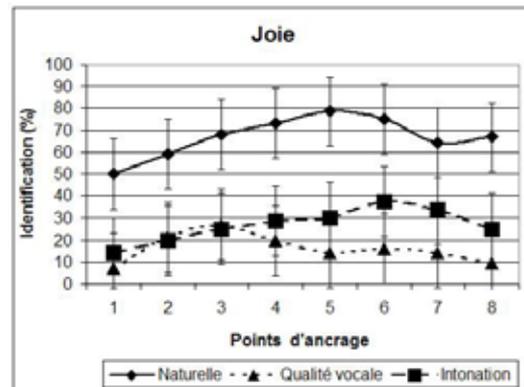


Figure 3 : Joie. Version 1 « naturelle », version 2 « qualité vocale », version 3 « intonation ».

Les résultats pour la *tristesse* mettent en lumière un fait intéressant : l'identification de la version 2 « qualité vocale » est aussi bonne que celle de la version naturelle, l'intonation restant proche du "zéro".

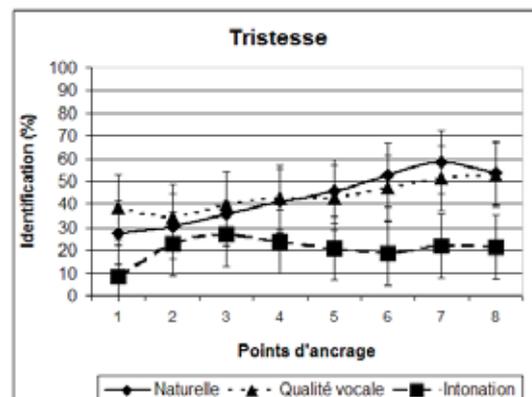


Figure 4 : Tristesse. Version 1 « naturelle », version 2 « qualité vocale », version 3 « intonation ».

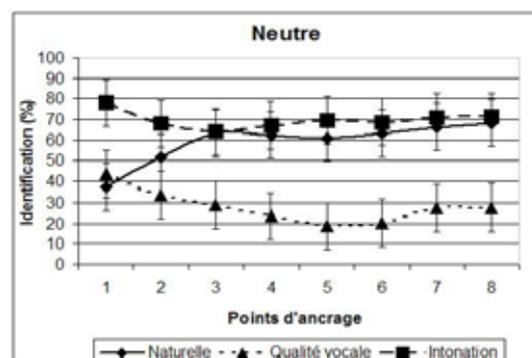


Figure 5 : Neutre. Version 1 « naturelle », version 2 « qualité vocale », version 3 « intonation ».

L'identification du neutre est équivalente pour la version « naturelle » et la version « intonation ». La participation de la qualité vocale n'est pas significative. Le neutre se distingue bien des modalités émotionnelles, surtout par l'intonation.

4. DISCUSSION

Le choix méthodologique d'utiliser le même énoncé pourvu d'un sens global purement référentiel pour une étude contextualisée de l'encodage des émotions a permis de déterminer de façon plus fiable la valeur perceptive de l'intonation et de la qualité vocale. Une telle méthode, en effet, nous a permis d'éviter deux biais. Le premier consiste à analyser des énoncés non écologiques dépourvus de signification, dans lesquels l'absence de structure syntaxique réduit la valeur perceptive de l'intonation. Le second, à l'inverse, repose sur l'analyse d'énoncés déjà très connotés par leur matériel verbal. Si pour ces derniers, on a pu montrer que la seule présence du lexique ne suffisait pas pour une identification optimale des émotions [4, 5], il restait très difficile de quantifier objectivement l'apport des paramètres suprasegmentaux dans les processus de reconnaissance.

La méthode présentée ici conforte l'hypothèse qui est la nôtre concernant le double statut de l'intonation dans l'activité de langage : 1) elle met en relief les mots informationnellement importants dans un contexte émotionnel spécifique; 2) elle facilite l'identification d'une émotion par un contour prosodique qui lui est propre. L'utilisation d'un énoncé lexicalement neutre permet d'évaluer le rôle du contour prosodique dans l'identification de l'émotion en l'isolant de l'amplification. Il est ensuite possible de caractériser ses propriétés acoustiques sur le plan qualitatif et quantitatif.

5. CONCLUSION

L'utilisation des patrons intonatifs naturels par la synthèse vocale donne accès, au moins partiellement, à l'expression des états affectifs. Certes, la qualité vocale joue un rôle généralement important, mais c'est un paramètre difficile à faire varier en temps réel. Le contour prosodique (hauteur, intensité, durée) est beaucoup plus facile à implémenter. Nos résultats montrent que le contour prosodique joue un rôle souvent non négligeable dans l'expression globale des émotions. L'analyse de la variabilité inter-individuelle du corpus et la sélection des stratégies les plus adaptées pour la reproduction dans la synthèse vocale constituent l'étape suivante de notre projet.

6. APPENDICE

Chagrin : « Tu sais comme j'aimais mon chien ? Hier, quand je suis revenu(e) de voyage, j'ai appris qu'il était mort. Je suis bouleversé(e). *Je vais rentrer à la maison maintenant.* J'vais jamais pouvoir le dire aux enfants. »

Colère : « Vous appelez ça une chambre d'hôtel ? Regardez un peu ces draps ! Ils sont ignobles. Vous ne croyez quand même pas que je vais dormir ici ! C'est révoltant ! *Je vais rentrer à la maison maintenant !* Ce n'est pas un hôtel ici, c'est un élevage de cafards ! »

Joie : « Mon frère reviendra demain ! Quelle joie ! Je suis si content(e) ! *Je vais rentrer à la maison maintenant !* Je vais annoncer cette super nouvelle à ma famille ! »

Neutre : « J'ai fini de ranger les boîtes. Elles sont classées et numérotées. *Je vais rentrer à la maison maintenant.* Je reviendrai demain à dix heures. »

Tristesse : « Ce que tu m'as appris m'a fichu le moral à zéro. C'est vraiment déprimant... *Je vais rentrer à la maison maintenant.* J'ai l'impression que c'est une situation sans issue. »

7. REMERCIEMENTS

Le projet était financé par le Conseil Régional de Basse-Normandie. Les auteurs remercient tous les participants aux expériences.

BIBLIOGRAPHIE

- [1] T. Bänziger and K.R. Scherer. The role of intonation in emotional expressions. *Speech Communication*, vol. 46, pages 252-267, 2005.
- [2] I. Yanushevskaya, C. Gobl and A. Ni Chasaide. Mapping voice to affect: Japanese listeners. In *Speech Prosody 2006*, paper OS4-4-265, 2006.
- [3] C. d'Alessandro. Voice source parameters and prosodic analysis. In: S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzy, I. Mleinek, N. Richter & J. Schließer (eds): *Methods in Empirical Prosody Research*. Berlin, New York: De Gruyter, pages 63-87, 2006.
- [4] I. Grichkovtsova, A. Lacheret and M. Morel. The role of intonation and voice quality in the affective speech perception. In *Proc. of Interspeech*, 2007.
- [5] I. Grichkovtsova, A. Lacheret, M. Morel, V. Beaucousin and N. Tzourio-Mazoyer. Affective speech gating. In *Proc. of ICPHS*, 2007.
- [6] I. Grichkovtsova. A cross-linguistic study of affective prosody production by monolingual and bilingual children: Scottish English and French. PhD thesis, Queen Margaret University, 2007.
- [7] A. Ghio, C. André, B. Teston and C. Cavé. PERCEVAL: une station automatisée de tests de PERCEPTION et d'EVALUATION auditive et visuelle. In *TIPA*, vol. 22, pages 115-133, 2007.
- [8] F. Grosjean. Gating. *Language and Cognitive Processes*, vol. 11, pages 597-604, 1996.
- [9] M.-N. Garcia, C. d'Alessandro, G. Bailly, P. Boula de Mareuil and M. Morel. A joint prosody evaluation of French text-to-speech systems. In *Proc. LREC*, pages 307-310, 2006.
- [10] M. Morel and A. Lacheret-Dujour. Kali, synthèse vocale à partir du texte : de la conception à la mise en oeuvre. In *TAL*, vol. 42, pages 1-29, 2001.