Dimensions acoustiques de la parole expressive : poids relatifs des paramètres resynthétisés par Praat vs. LF-ARX

Nicolas Audibert[1], Damien Vincent[2], Véronique Aubergé[1], Albert Rilliard[1] & Olivier Rosec[2]

[1] Institut de la Communication Parlée
CNRS UMR 5009, Grenoble, France
[2] France Telecom, R&D Division
{audibert, auberge, rillard}@icp.inpg.fr, {damien.vincent, olivier.rosec}@francetelecom.com

ABSTRACT

The emotional prosody is multi-dimensional. A debated question is whether some parameters are more specialized to convey some emotion dimensions. Selected stimuli carrying acted expressions of anxiety, disappointment, disgust, disquiet, joy, resignation, satisfaction and sadness on monosyllabic words were used to synthesize artefactual stimuli by projecting separately prosodic parameters on neutral expressions, with Praat and an LF-ARX algorithm. Perceptive evaluation of stimuli and comparison of results (1) indicate that F0 contours bring more information on positive expressions while voice quality and duration convey more information on negative expressions, and intensity alone is not informative enough (2) diagnoses minor artifacts of both synthesis methods which consequences may have interesting implications in expressive speech synthesis (3) validates the efficiency of the LF-ARX algorithm (4) measures the relative weights of each of the LF-ARX voice quality parameters.

1. INTRODUCTION

La multi-dimensionnalité de la prosodie des affects est un problème complexe, le débat restant ouvert quant à la spécificité de certains indices émotionnels pour l'expression d'affects particuliers. Une autre question non résolue est celle de la description de la qualité de voix comme une seule ou plusieurs dimensions. Afin d'évaluer le poids de chaque paramètre prosodique dans la morphologie vocale des émotions, et donc dans la perception de leurs expressions, il est nécessaire d'évaluer comment des stimuli porteurs de ces expressions sont perçus lorsqu'on agit séparément sur ces paramètres. Étant donné que les variations de ces paramètres résultent d'un contrôle global du conduit vocal, il semble toutefois impossible de recueillir naturellement de tels stimuli, quand bien même des locuteurs seraient spécifiquement entraînés pour cette tâche.

Nous avons donc adopté une méthode basée sur la resynthèse : une analyse acoustique de stimuli de référence est effectuée avant de synthétiser de nouveaux stimuli à partir de tout ou partie des paramètres analysés, en fonction des hypothèses à tester ; une évaluation perceptive de ces stimuli permet ensuite de valider ces hypothèses. Une telle méthode a déjà été utilisée dans une série d'études relatives à l'expression des affects dans la parole. Ainsi, le rôle de la qualité de voix dans les expressions émotionnelles a été attesté à l'aide de stimuli resynthétisés après modification de l'onde de débit glottique analysée sur des stimuli de référence [8]. Par ailleurs [5], les variations de F0 et de durée extraites de diverses expressions émotionnelles ont été appliquées en synthèse par concaténation à des diphones porteurs d'expressions émotionnelles différentes. L'évaluation perceptive des stimuli ainsi construits a conduit les auteurs à conclure que les expressions de colère étaient majoritairement véhiculées par les diphones, la tristesse par F0 et la durée, tandis qu'aucune conclusion claire ne pouvait être tirée quant aux expressions de joie. Enfin, dans un certain nombre d'études (par ex. [3]), des stimuli ont été synthétisés à partir de mesures multiparamétriques afin d'évaluer la pertinence des paramètres acoustiques mesurés pour la perception d'expressions émotionnelles.

Deux méthodes ont été successivement utilisées pour réaliser cette resynthèse paramètre par paramètre, les stimuli générés étant ensuite évalués perceptivement selon le même protocole. La première (présentée en section 2), basée sur Praat [4], ne permet pas de manipuler directement la qualité de voix. En revanche la seconde (présentée en section 3), basée sur un modèle ARX [6] excité par une source LF [7], permet d'évaluer séparément le rôle des différents paramètres de qualité de voix.

2. ETUDE 1 : RESYNTHÈSE PRAAT

2.1. Choix des données de référence

10 stimuli extraits du corpus E-Wiz / Sound Teacher [1] et constituant un sous-ensemble des 72 stimuli actés précédemment évalués dans une étude perceptive [10] ont été sélectionnés. Ces stimuli expriment sur les mots monosyllabiques [ʁuʒ] et [sabl] les états émotionnels suivants, joués par un acteur après avoir été ressentis dans une tâche « piège » prétextant une aide à l'apprentissage phonétique des langues étrangères : anxiété, déception, dégoût, inquiétude, joie, résignation, satisfaction, tristesse et neutre. La pertinence d'une expression neutre, sélectionnée comme référence pour la comparaison des contours mutiparamétriques, est validée par la présence dans les productions spontanées de ce locuteur d'expressions étiquetées par lui-même comme « rien ».

2.2. Méthode de synthèse

L'analyse des contours multiparamétriques et la synthèse ont été effectuées à l'aide la fonction de synthèse basée sur TD-PSOLA de Praat [4], selon une procédure semi-

automatique consistant à styliser les contours de F0 et d'intensité extraits d'un stimulus source puis à appliquer tout ou partie de ces contours à un stimulus cible et générer un nouveau stimulus combinant des propriétés des stimuli source et cible. La stylisation et la transplantation des contours ont été contrôlées afin d'éviter de négliger des points saillants. La transplantation des contours d'intensité a été réalisée en appliquant un contour relatif puis en rééchelonnant le signal pour obtenir la même valeur globale d'énergie.

2.3. Stimuli générés

Pour chacun des stimuli originaux porteurs d'une expression émotionnelle, 5 stimuli distincts ont été générés, étiquetés en fonction des paramètres du stimulus cible utilisés : (i) *contrôle*, construit en appliquant les contours stylisés de F0 et d'intensité du stimulus source à lui-même et destiné à évaluer d'éventuels artefacts dus au processus de resynthèse (ii) *F0*, construit en appliquant le contour stylisé de F0 du stimulus source au stimulus porteur d'une expression neutre correspondant au même mot (iii) *intensité,* obtenu en appliquant le contour d'intensité du stimulus source à l'expression neutre correspondante (iv) *F0 et intensité,* construit en appliquant les contours de F0 et d'intensité à l'expression neutre correspondante (v) *QV et durée*. Cette dernière condition a été obtenue en appliquant les contours de F0 et d'intensité de l'expression neutre au stimulus source. Ainsi seuls les phénomènes de durée et de qualité de voix (QV) du stimulus source subsistent, tandis que ses variations spécifiques de F0 et d'intensité sont neutralisées. En complément des 40 stimuli générés à partir des 8 expressions émotionnelles sélectionnées, un stimulus en condition *resynthèse complète* a été généré pour chacune des 2 expressions neutres.

2.4. Evaluation perceptive

Les 42 stimuli générés ont été notés par 40 juges de langue maternelle française (6 hommes, 34 femmes, d'âge moyen 23,3 ans) en chambre sourde, en ordre aléatoire, avec 3 présentations non consécutives de chaque stimulus La présentation des stimuli et l'enregistrement des réponses ont été automatisés à l'aide d'une interface. Les sujets avaient pour instruction de sélectionner l'une des 8 expressions proposées (anxiété, déception, dégoût, inquiétude, joie, résignation, satisfaction ou tristesse) ou l'étiquette neutre. De plus il leur était demandé de noter l'intensité émotionnelle perçue entre 1 et 10.

2.5. Résultats

La valeur élevée de l'alpha de Cronbach (α=.95) indique que les réponses données par les différents juges sont cohérentes. Les résultats ont été traités séparément pour chaque condition de resynthèse. Etant donné que les intensités émotionnelles sont significativement corrélées au nombre de réponses attribuées aux différentes étiquettes (r^2=.854) et apportent peu d'information supplémentaire, seuls les résultats relatifs aux scores d'identification sont discutés ici. Afin de prendre en compte les principales confusions faites par les sujets en condition de contrôle, similaires à celles observées lors de la validation perceptive des stimuli originaux [10], et faire ressortir plus clairement les principales tendances, certaines étiquettes ont été regroupées : joie avec satisfaction, anxiété avec inquiétude, tristesse avec déception et résignation, tandis que dégoût et neutre demeurent des catégories distinctes. Le test du khi-deux indique que les distributions après regroupements sont significativement différentes du hasard (p=.01). Une fois ces regroupements effectués, l'essentiel des informations pertinentes se trouve sur les diagonales des matrices de confusion, qui correspondent aux scores d'identification. Les données après regroupement ont donc été converties en bonnes ou mauvaises réponses et normalisées pour permettre une évaluation statistique des différences observées au moyen d'une ANOVA sur mesures répétées.

Une première observation est que les scores obtenus en conditions « manipulées » sont moins élevés qu'en condition de contrôle, indiquant qu'aucune dimension ne véhicule seule toute l'information affective. Néanmoins cette différence n'est pas significative en condition *QV et durée* pour anxiété et inquiétude (47,2% vs. 55,6% en condition de contrôle) ni pour tristesse, résignation et déception (55,8% vs. 59,6%), la qualité de voix et la durée apparaissant comme véhiculant l'essentiel de l'information pour ces expressions négatives. Bien que ce soit également le cas pour le dégoût, les autres dimensions ne portant que très peu d'information, il est surprenant de constater que cette expression est significativement moins bien reconnue qu'en condition de contrôle (34,2% vs. 61,7%). En condition *intensité*, seules les expressions de tristesse, déception et résignation (identifiées à 35,6%) ont obtenu un score supérieur à 10%. Toutefois de nombreux sujets percevant l'une de ces expressions en condition *intensité* ont également confondu les expressions neutres avec tristesse, résignation ou déception en condition de contrôle. Ceci explique également le score plus élevé en condition *F0* (36,9%) qu'en condition *F0 & intensité* (26,7%) obtenu par ces expressions. En conditions *F0* et *F0 et intensité*, les expressions de joie et satisfaction ont été les mieux reconnues (respectivement à 56,3% et 67,5% vs. 85,4% en condition de contrôle). De plus, bien que l'intensité soit insuffisante seule, la comparaison des scores en conditions *F0* et *F0 et intensité* montre qu'elle apporte un gain significatif (p=.05) pour les expressions de joie et satisfaction (56,3% vs. 67,5%) ainsi que d'anxiété et d'inquiétude (26,7% vs. 21,3%), et qu'elle ne peut donc pas être considérée comme non informative.

3. ETUDE 2 : RESYNTHÈSE LF-ARX

3.1. Méthode de synthèse

Le modèle LF de production de la parole [7], sur lequel la méthode de synthèse utilisée ici est basée, s'inscrit dans une approche source-filtre : la source correspond à l'excitation glottique modifiée pour y intégrer l'effet des lèvres (dérivateur), ou onde de débit glottique dérivée, et le filtre modélisé aux résonances du conduit vocal. Dans le cadre de la modélisation d'un son voisé par un modèle ARX [6]

excité par une source LF, l'analyse revient à estimer les 3 paramètres du modèle LF décrivant la source, une composante stochastique appelée résidu, les coefficients du filtre correspondant au conduit vocal, la fréquence fondamentale et l'énergie. Etant donné que les paramètres du filtre et du résidu peuvent être estimés par la méthode des moindres carrés une fois les paramètres LF connus, une méthode efficace basée sur une recherche exhaustive dans un espace d'ondes LF quantifiées a été proposée pour l'estimation de ces paramètres [11].

Dans ce cadre, le processus de resynthèse consiste à remplacer certains paramètres issus de l'analyse du stimulus neutre (source) par les valeurs de ces paramètres obtenues par l'analyse du stimulus « émotionnel » (cible). Ce processus comprend une procédure d'alignement ainsi qu'un algorithme de synthèse. L'alignement nécessite que les stimuli source et cible aient le même contenu phonétique : après avoir apparié les frontières phonémiques des 2 stimuli, les points analysés dans un même phonème sont reliés par un mécanisme d'interpolation linéaire. Le résultat de cet alignement doit être contrôlé car des erreurs peuvent apparaître en cas de non congruence entre les informations de voisement des stimuli source et cible. L'algorithme de synthèse, similaire à ceux utilisés pour les modifications prosodiques basées sur TD-PSOLA [9], détermine les instants de synthèse et génère pour chacun de ces instants une paire de trames issues respectivement des stimuli source et cible, la suite du processus de synthèse devenant alors triviale.

3.2. Stimuli générés

Les 10 stimuli de référence de l'étude 1 ont été à nouveau utilisés comme base pour la resynthèse par l'algorithme LF-ARX. Toutefois la méthode de synthèse utilisée n'a pas permis de générer des stimuli de qualité suffisante à partir de l'expression de la satisfaction, qui a donc dû être éliminée de cet ensemble. Seules 7 conditions de synthèse ont été retenues parmi les combinaisons possibles des 6 jeux de paramètres (la qualité de voix étant considérée comme décrite par l'ensemble source, résidu et filtre), étiquetés en fonction des paramètres du stimulus expressif utilisés : (i) *contrôle* (ii) *QV et durée* (iii) *QV* (iv) *source et résidu* (v) *source*, (vi) *durée* et (vii) *F0 et intensité*. Les conditions *contrôle*, *QV et durée* et *F0 et intensité* permettent une comparaison directe avec les résultats de l'étude 1.

3.3. Evaluation perceptive et résultats

Les 51 stimuli générés ont été évalués par 25 juges de langue maternelle française (7 hommes, 18 femmes, d'âge moyen 25,7 ans), selon le même protocole que celui utilisé dans l'étude 1.

Les résultats de cette évaluation perceptive présentent également une valeur élevée d'alpha de Cronbach (α=.92), ainsi qu'une corrélation significative entre les intensités émotionnelles perçues et le nombre de réponses attribuées aux étiquettes correspondantes (r^2=.889). De plus les confusions observées en condition de contrôle étant similaires à celles de l'étude 1, les mêmes regroupements ont été appliqués. L'expression de la satisfaction n'étant pas présente dans les stimuli utilisés ici, l'expression de la joie a été traitée comme une catégorie distincte. De même que dans l'étude 1, les données ont été converties en bonnes ou mauvaises réponses et normalisées pour tester la significativité (p=.01) des différences observées par des analyses de variance sur mesures répétées.

Les principaux résultats de l'étude 1 sont ici confirmés : la joie est significativement mieux reconnue que les autres expressions en condition *F0 et intensité*, quoique avec un score significativement moins élevé qu'en condition de contrôle (58,7% vs. 77,3%) ; les expressions de tristesse, résignation et déception, de même que celles d'anxiété et d'inquiétude ne sont pas significativement moins bien reconnues en condition *QV et durée* qu'en condition de contrôle (respectivement 52,9% vs. 56% et 60% vs. 67,3%). On retrouve également pour l'expression du dégoût le même phénomène que dans l'étude 1, cette expression étant significativement moins bien reconnue en condition *QV & durée* qu'en condition de contrôle (42,7% vs. 70,7%), alors que F0 et intensité portent très peu d'information (1,3%). Des observations peuvent en outre être tirées des conditions de synthèse absentes de la première évaluation, les dimensions testées conjointement sous l'étiquette *QV & durée* ayant ici fait l'objet d'une évaluation séparée. Ainsi en condition *durée* les expressions de tristesse, résignation et déception ont été aussi bien reconnues qu'en condition de contrôle (56,9% vs. 56%), tandis que ces expressions ont été significativement moins bien reconnues en condition *QV*, mais néanmoins au dessus du hasard (44,4%). Les expressions d'anxiété et d'inquiétude, quant à elles, présentent la même quantité d'information affective en condition *QV* et en condition *durée* (identifiées à 46% dans ces 2 conditions). Enfin, la majeure partie de l'information affective pour le dégoût est portée par la durée (49,3% vs. 8% en condition *QV* et 1,3% en condition *F0 et intensité*). D'autre part la comparaison des scores obtenus en conditions *VQ*, *source & résidu* et *source* permet d'évaluer les influences relatives des différents paramètres de modélisation de la qualité de voix. Si la source apparaît comme porteuse de toute l'information de qualité de voix pour les expressions de tristesse, résignation et déception ainsi que pour la joie (pas de gain significatif en ajoutant le filtre et le résidu), les expressions d'anxiété et d'inquiétude ont été significativement mieux reconnues en condition *source & résidu* (35,3%) qu'en condition *source* (24%), et significativement mieux en condition *QV* (46%) qu'en condition *source & résidu*.

Etant donné les scores obtenus pour l'expression de la joie en condition *QV* (10,7%) et en condition *durée* (0%), on devrait également obtenir un score d'environ 10% pour cette expression en condition *QV & durée*. Or ce score est largement supérieur à la valeur attendue (30,7%), ce qui nous a alerté sur la possible présence d'un artefact. Un examen attentif des signaux a révélé la présence d'un bruit de fermeture très court et d'énergie moyenne au début du signal généré en condition *QV & durée*, qui peut avoir été

interprété par les juges comme un coup de glotte annonciateur d'un rire, rendant le score d'identification correspondant artificiellement élevé. En effet l'algorithme de synthèse, étalonné sur des signaux dans lesquels les portions étiquetées comme silence sont effectivement silencieuses, génère les segments silencieux en copiant les parties étiquetés comme silence du stimulus source ou cible. Ce bruit étant présent dans l'expression neutre utilisée, il a donc été automatiquement copié au début du signal généré dans cette condition.

4. DISCUSSION

La table 1 présente les scores obtenus après regroupement dans les études 1 et 2 (étiquetés respectivement *Praat* et *ARX*) pour les conditions de synthèse communes. Afin de rendre les comparaisons possibles, les confusions des expressions de joie avec la satisfaction ont ici été prises en compte dans le calcul des scores d'identification de la joie dans l'étude 1, d'où la différence entre certains scores présentés dans cette table et ceux présentés en section 2.

Table 1 : Scores d'identification obtenus dans les études 1 et 2 après regroupement.

		contrôle	F0+int	QV+dur
joie	Praat	70,8%	42,5%	6,7%
	ARX	77,3%	58,7%	30,7%
trist, res, décep	Praat	59,6%	26,7%	55,8%
	ARX	56%	27,1%	52,9%
anxiété, inq,	Praat	55,6%	21,4%	47,2%
	ARX	67,3%	40,7%	60%
dégoût	Praat	61,7%	3,3%	34,2%
	ARX	70,7%	1,3%	42,7%
neutre	Praat	31,7%		
	ARX	52,7%		

Les scores obtenus en condition de contrôle sont généralement plus élevés avec la synthèse LF-ARX qu'avec la synthèse Praat, à l'exception des expressions de tristesse, résignation et déception pour lesquelles cette différence est faible (la structure des données ne permet pas de tester la significativité des différences entre les résultats des études 1 et 2). Le codage des expressions par l'algorithme LF-ARX semble donc globalement meilleur qu'avec Praat.

Par ailleurs, si on considère qu'en l'absence de l'artefact décrit en section 3 le score de 30,7% obtenu par l'expression de la joie en condition *QV & durée* dans l'étude 2 aurait dû être proche de 10%, on peut s'interroger sur la différence entre ce score théorique et celui obtenu dans l'étude 1 (6,7%). Un artefact de la synthèse Praat pourrait être responsable de ce score plus faible : la méthode de transplantation des contours d'intensité contrôlant la valeur globale d'énergie mais non les valeurs locales, l'intensité à la fin du stimulus généré est plus faible que celle à la fin de l'expression neutre. Cette intensité finale peu élevée peut donc avoir été interprétée par les juges comme incompatible avec une expression de joie, ce qui poserait alors la question de la définition d'un contour d'intensité.

Les résultats de ces deux études montrent donc que l'information affective des expressions positives est principalement véhiculée par les contours de F0, tandis qu'elle est surtout portée par la durée, et dans une moindre mesure par la qualité de voix pour les expressions négatives. Ils valident également la qualité du codage réalisé par l'algorithme LF-ARX, de même que la pertinence de la modélisation des informations de filtre et de résidu. Les conséquences des artefacts observés ont d'intéressantes implications potentielles en synthèse, puisqu'un phénomène local et aussi mineur en termes de quantité d'information peut modifier la perception des affects exprimés.

BIBLIOGRAPHIE

[1] V. Aubergé, N. Audibert and A. Rilliard. E-Wiz: A Trapper Protocol for Hunting the Expressive Speech Corpora in Lab. *4th LREC*, Lisbonne, pages 179-182, 2004.

[2] V. Aubergé, N. Audibert and A. Rilliard. Acoustic Morphology of Expressive Speech: What about Contours? *Speech Prosody 2004*, Nara, pages 201-204, 2004.

[3] T. Bänziger, M. Morel and K. R. Scherer. Is there an emotion signature in intonational patterns? And can it be used in synthesis? *Eurospeech 2003*, Genève, pages 1641-1644, 2003.

[4] P. Boersma and D. Weenink. Praat: doing phonetics by computer. http://www.praat.org.

[5] M. Bulut, S. Narayanan and A. Syrdal. Expressive speech synthesis using a concatenative synthesizer. *7th ICSLP*, Denver, pages 1265–1268, 2002.

[6] W. Ding, H. Kasuya and S. Adachi. Simultaneous estimation of vocal tract and voice source parameters based on an ARX model. in *IEICE Trans. Inf. Syst.*, E78-D (6), pages 738–743, 1995.

[7] G. Fant, J. Liljencrants and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR* (4), pages 1–13, 1985.

[8] C. Gobl and A. Ní Chasaide. The role of the voice quality in communicating emotions, mood and attitude *Speech Comm.* (40), pages 189–212, 2003.

[9] E. Moulines and J. Laroche. Non-parametric techniques for pitch-scale and time-scale modifications of speech. *Speech Comm.* (16), pages 175–205, 1995.

[10] A. Rilliard, V. Aubergé and N. Audibert. Evaluating an Authentic Audio-Visual Expressive Speech Corpus. *4th LREC*, Lisbonne, pages 175-178, 2004.

[11] D. Vincent, O. Rosec and T. Chonavel. Estimation of LF glottal source parameters based on arx model. *Interspeech 2005*, Lisbonne, pages 333-336, 2005.