Génération de la prosodie par superposition de contours chevauchants: application à l'énonciation de formules mathématiques

Bleicke Holm & Gérard Bailly

Institut de la Communication Parlée - UMR CNRS n^o5009 - INPG & Université Stendhal
46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France
Tél: ++33 04 76 57 47 11 - Fax: ++33 04 76 57 47 10
e-mail: holm@icp.inpg.fr - http://www.icp.inpg.fr/

Abstract

We present here a model for generating prosody by superposing overlaping multiparametric contours. These contours are associated with high-level communication tasks such as segmentation, hierarchisation or emphasis of units. We propose a scheme for automatically learning these contours and apply this new paradigm to the enunciation of mathematical formulae.

1. Introduction

La prosodie participe à la transmission d'information linguistique dans l'acte de parole. Parmi les fonctions linguistiques assumées les plus étudiées, on trouve la fonction de hiérarchisation/segmentation d'unités. Cutler note ainsi : "*Two of the major uses of prosodic information in situations of communication are to encode salience and segmentation.*" [6, p.264]. Si tout le monde s'accorde maintenant sur le rôle fondamental joué par la prosodie dans l'acquisition [14] et la structuration du langage, de grosses divergences apparaissent sur les moyens d'encodage de l'information. Dans le cadre d'un modèle de génération automatique de la prosodie, il s'agit de fournir une évolution acceptable de paramètres prosodiques en fonction de la structuration de l'énoncé. Se pose ainsi la double question du Quoi et du Comment : quelles sont les informations véhiculées et comment sont-elles encodées ?

L'encodage d'informations linguistiques, supposées discrètes et dénombrables, par un continuum a été décrit par de multiples approches. La plupart utilisent une interface phonologique autonome où la structuration linguistique est "traduite" en une structuration phonologique organisant les faits prosodiques saillants en des structures tonales et accentuelles. A cette étape de transfert phonologique, doit alors succéder une étape de génération phonétique où les structures phonologiques sont traduites en continuums prosodiques.

Les systèmes de génération automatique différent dans la richesse des structures phonologiques exploitées : ainsi, Black et Hunt [5] utilisent la transcription ToBI [17], Di Cristo et al IntSyn [10]. D'autres auteurs utilisent une description plus économique de la structure phonologique en termes de groupes accentuels [18, 12, 13]. Ce dernier groupe d'approches est très hétérogène : Malfrère et al [12] portent leur effort sur une concaténation élégante de contours faiblement étiquetés par rapport à leur fonction alors que Pasdeloup [16] focalise sur un bon découpage en groupes accentuels.

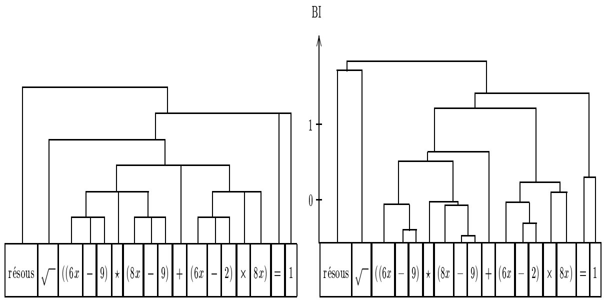
Les modèles phonétiques utilisés lors de l'étape de génération vont de la simple connection de points cibles pour le génération de la mélodie [10] à des modèles plus complexes. La notion de superposition est largement répandue au moins ce qui concerne deux composants qui sont distingués : phrase et groupe accentuel. La superposition peut être très explicite, comme dans [13] ou en forme de l'imposition d'une ligne de déclinaison sur laquelle s'ajoutent les contributions dues aux accents. Même [5] incorpore une superposition car la réalisation des accents est exprimée par rapport à une fréquence de référence qui varie au cours de la phrase. Chez Traber [18], l'usage d'une large fenêtre sur l'énoncé et de réseaux récurrents permet de compenser la pauvreté de la description phonologique d'entrée : le réseau peut "calculer" de manière implicite des niveaux de représentation intermédiaire.

Notons finalement la proposition d'Aubergé [1] qui contraste avec cette décomposition des tâches : c'est le niveau d'encodage phonétique qui traduit "directement" la structuration linguistique par superposition de contours globaux coextensifs aux unités qu'ils qualifient. Dans cet article, nous présentons un encodage à la fois *simple* par le mécanisme d'encodage qu'il utilise et *riche* par les capacités de structuration et de coarticulation qu'il offre. L'encodage opère par superposition de contours chevauchants. Cet encodage est *simple* et général car chaque contour code une fonction de hiérarchisation/segmentation d'unités quelque soit cette unité et sa hauteur dans la hiérarchie linguistique. *Riche* parce que le nombre de niveaux superposés et donc le degré de hiérarchisation de l'énoncé est illimité.

2. L'énonciation de formules mathématiques

Notre corpus a été conçu afin de pouvoir étudier la manière avec laquelle la prosodie peut encoder des relations de dépendance à l'intérieur d'une phrase. Nous avons choisi des formules mathématiques lues (FM) parce qu'on y trouve une structure syntaxique profonde qui est souvent ambiguë à l'oral et qui nécessite un recours renforcé à la structuration de l'énoncé par la prosodie. Nous avons restreint le domaine des formules aux équations algébriques issues de l'enseignement de niveau 3^e. Le corpus a été généré automatiquement en variant systématiquement la profondeur syntaxique et la longueur syllabique des constituants pour couvrir statistiquement l'influence de ces paramètres sur la stratégie prosodique – tout en gardant une taille de corpus raisonnable. Nous obtenons un corpus de 157 FM[1]. Il est lu par un locuteur français à qui nous avons donné la consigne de ne pas utiliser d'indicateurs lexicaux de structure comme "ouvrez les parenthèses".

Ainsi que montré sur la figure 1 et plus amplement commenté dans [11], la structuration prosodique de formules énoncées est riche et profonde. De plus, chaque ajout d'un niveau syntaxique supplémentaire se traduit par un accroissement de la hauteur de la structure de performance correspondante, ce qui est bien en adéquation avec le modèle de superposition dont la description suit.



(a) Syntaxe (b) Performance

Fig. 1: Structure syntaxique et structure de performance [8] de l'énoncé d'une formule.

3. Le modèle

Dans la proposition initiale d'Aubergé [1], chaque niveau linguistique génère en fonction de sa nature, de sa fonction et de sa longueur un ensemble de contours multiparamétriques qui seront ensuite concaténés et superposés. Morlec [15] a traité le problème de l'apprentissage des contours, de leur expansion avec la taille des unités et de leur enchaînement en proposant une architecture connexionniste modulaire. Cette approche a été validée aussi bien en analyse qu'en synthèse.

Ce modèle initial et son implémentation souffrent cependant de deux limitations : (a) le modèle suit de trop près la syntaxe au sens où un contour est indexé par des caractéristiques syntaxiques souvent redondantes et (b) si l'implémentation par réseau récurrent gère mieux la coarticulation que la simple concaténation initiale - où l'étiquetage en sous-groupes était conditionné à des contraintes phonotactiques -, elle fait perdre la notion initiale de lexique de contours attachés à une fonction linguistique.

Le modèle que nous développons ici vise à résoudre ces deux problèmes par (a) l'adoption d'une hypothèse additionnelle donnant au modèle morphologique initial une véritable morphophonologie fonctionnelle, et (b) l'adoption d'un modèle superpositionnel de contours *chevauchants*.

3.1. Une morphophonologie fonctionnelle

La morphophonologie fonctionnelle que nous proposons suppose que l'intonation met au service de la syntaxe des fonctions générales, que l'intonation partage avec la syntaxe mais qu'elles assument de manière autonome et donc pas forcément de manière congruente. Parmi ces fonctions plus autonomes, on peut citer les fonctions attitudinales [15] ou émotionnelles que l'on peut caractériser par des formes prosodiques prototypiques. Parmi les fonctions plus synergétiques, on peut citer la fonction de segmentation/hiérarchisation ou de mise en relief. Dans ce papier, nous allons poser et tester l'hypothèse que la prosodie possède les moyens de segmenter/hiérarchiser des unités consécutives en centrant un contour sur la joncture entre ces deux unités (voir figure 2(en haut)). La hiérarchie est alors encodée par la superposition de contours qui se chevauchent et qui ont des empans différents mais limités aux unités adjacentes à la joncture mise en valeur.

L'encodage par contours permet de distribuer l'information au cours du temps en ne limitant pas l'accès à la fonction au décodage d'éléments saillants du discours. Cet encodage distribué permet d'expliquer les phénomènes d'anticipation [2], d'attente et de prédiction sur la suite d'un énoncé [9]. Il semble cependant évident que cet encodage doive être programmé sur un empan temporel limité.

3.2. Apprentissage des contours

L'inversion d'un tel modèle est un problème malposé : comment obtenir les contours localisés à partir d'un corpus qui ne contiendra jamais ces contours isolés mais toujours superposés ? La stratégie adoptée par Aubergé [1], reprise par Morlec [15], consiste à apprendre hiérarchiquement les contours, de celui possédant l'empan le plus large jusqu'au plus petit. A la différence de Morlec où chaque module génère des contours *globaux* associés à un niveau hiérarchique *fixe*, chaque module ici prend en charge la génération de tout contour localisé assumant la fonction désirée : le module génère des contours ancrés sur une partie du discours et non plus sur le discours entier. Il est appliqué plusieurs fois aussi bien de manière séquentielle (des fonctions identiques apparaissant à divers points du discours) que hiérarchique (appliqué à des niveaux linguistiques différents pour par exemple segmenter/hiérarchiser des unités de différentes tailles). La décomposition du contour global du discours en contours localisés et ancrés ne peut donc suivre l'apprentissage hiérarchique.

Nous proposons un apprentissage itératif permettant à chaque module de s'approprier petit à petit des parties du contour global. Chaque cycle d'apprentissage comporte deux phases : (1a) calcul de la différence entre contours globaux observés et prédits, (1b) calcul des cibles (contours localisés) de chaque module

[1]Pour plus de précisions sur le corpus on peut se référer à [11].

et, (2) apprentissage des cibles par les modules (RN). (1a) est triviale – sauf au premier cycle où nous utilisons une prédiction égale à zéro. La base pour le calcul des nouveaux cibles en (1b) sont les contours localisés prédits. On y ajoute de manière uniforme une fraction de la différence calculée en (1a) : pour chaque trame[2], chaque contour localisé contribuant au contour prosodique de cette trame reçoit la différence en ce point divisée par le nombre total de contours localisés contributeurs – en s'assurant ainsi que la superposition des cibles est égale au contour observé. La phase (2) s'appuie sur l'implémentation des modules en forme de réseaux neuronaux : ils sont bien adaptés pour mettre en cohérence les nombreux exemplaires de réalisation d'un même contour localisé : chaque module génère une expansion du mouvement d'un contour prototypique (voir par exemple Figure 3), chaque réalisation ne différant que par l'empan temporel et son lieu d'ancrage dans cet empan.

3.3. L'énonciation de formules

La fonction de segmentation/hiérarchisation est la plus prégnante des fonctions sollicitées lors de l'énonciation de formules. La fonction de hiérarchisation est essentiellement sollicitée pour mettre en valeur l'enchâssement des opérateurs. Les opérateurs présents dans notre corpus sont simples : on distingue des opérateurs unaires (comme $\sqrt{\ldots}$) et des opérateurs binaires (comme +). Nous travaillerons prochainement sur des opérateurs plus complexes du type $\sum$.

Ainsi selon l'hypothèse que l'intonation suit ici la hiérarchie syntaxique, nous nous contraignons à distinguer dans notre corpus que deux types de dépendance : dépendance gauche (G) – entre un opérateur et son opérande[3] à gauche et, dépendance droite (D) entre un opérateur et son opérande à droite. À chaque frontière, entre deux feuilles de l'arbre syntaxique, nous posons un marqueur G ou D. Ils symbolisent chacun une relation de dépendance, à l'image des systèmes de marques syntaxiques proposées dans la littérature [7, 3]. Deux modules prendront dès lors en charge la génération de toutes les instances de ces marques en fonction des empans correspondants.

Les relations de présuppositions entre phrases ne sont pas factorisées avec les précédentes : nous avons ainsi un troisième module prenant en charge le contour ancré à la frontière entre "Résous :" et la formule. Son empan sera l'énoncé entier.

4. Premières observations

Afin de donner une idée de la qualité des contours générés, nous présentons dans la figure 2 pour un exemple la prédiction (trait plein) avec le contour observé (pointillés). On observe un comportement similaire mais, notamment l'accent sur la sixième syllabe n'est pas bien rendu.

Il est indispensable d'analyser le comportement

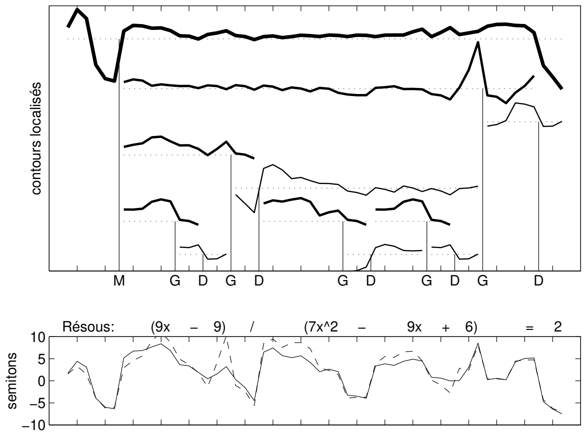


Fig. 2: En haut : contours localisés et ancrés de f_0. On note le point d'ancrage de chaque contour dont on distingue trois types : hiérarchie de phrase (M), dépendance gauche (G) et dépendance droite (D). Les pointillés indiquent le zéro pour chaque contour. — En bas : contour observé (tirets) contour prédit par superposition (plein)

des contours localisés générés. Regardons quelques contours du coefficient d'allongement que produit le module G (figure 3). Dans notre corpus, ces contours apparaissent toujours pour un opérateur et son opérande gauche. Dans la ligne en bas, on voit que la frontière est marquée par un important allongement[4] du GIPC avant l'opérateur. L'allongement décroît avec la longueur de l'unité gauche (en remontant les lignes dans la figure 3). La première ligne des contours échappe à cette régularité : un opérande d'une syllabe est clairement marqué par un fort allongement.

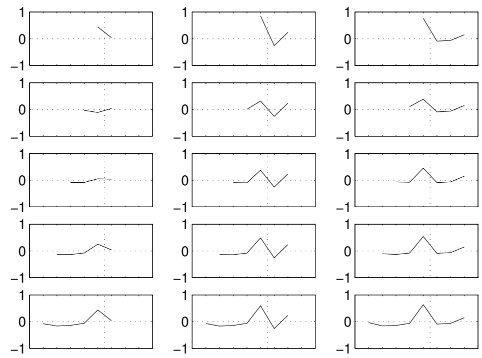


Fig. 3: Sortie (coefficient d'allongement) du module G pour différentes longueurs des unités : de 1 à 5 GIPC avant le point d'ancrage (indiqué par les lignes verticales) et de 1 à 3 après celui-ci.

5. Évaluation

En guise de première évaluation, nous présentons une comparaison des RMS-erreurs pour plusieurs méthodes de génération des contours : (1) "global–non-hiérarchique", (2) "global–hiérarchique" et (3) "local–

[2]L'unité de programmation prosodique est pour l'instant le GIPC [4] caractérisé par trois points de F0 et un coefficient d'allongement.

[3]Par "opérande" nous entendons l'ensemble des feuilles appartenant à un sous-arbre en direct voisinage de l'opérateur.

[4]Voire une pause. Pour le traitement des pauses dans notre modèle, on peut se référer à [4].

apprentissage	C1+C2	C1	
test	C1+C2	C1	C2
méth.(1)	0.205±0.010	0.185±0.005	0.248±0.003
méth.(2)	0.217±0.004	0.206±0.003	0.248±0.010
méth.(3)	0.231±0.031	0.210±0.005	0.239±0.008
méth.(1)	1.96±0.03	1.81±0.04	2.40±0.06
méth.(2)	2.07±0.01	1.98±0.02	2.34±0.04
méth.(3)	2.23±0.20	2.12±0.04	2.31±0.02

Tab. 1: Erreur de prédiction et son écart-type pour plusieurs méthodes et différents corpus d'apprentissage. En haut : coefficient d'allongement, en bas : f_0 en semitons.

itératif". Les deux premières méthodes suivent l'architecture de Morlec, mais diffèrent par la gestion de l'apprentissage : dans (1) l'apprentissage des modules se fait simultanément (s'approchant ainsi du modèle de Traber) tandis que dans (2) l'apprentissage est hiérarchique (comme proposé par Morlec). (3) correspond au modèle proposé dans cet article. Dans le tableau 1 la première colonne correspond à un apprentissage sur le corpus total (C1+C2). Les deux autres montrent les erreurs pour des apprentissages sur une moitié du corpus (C1) – au milieu : l'erreur sur le corpus d'apprentissage C1, et à gauche : celle sur celui de test (C2). Chaque valeur est issue de 4 apprentissages du même type. Bien que l'erreur totale augmente avec le numéro de version, on observe une amélioration pour la généralisation (colonne de droite) – ce qui est le critère plus important pour l'évaluation du modèle.

Notons finalement que pour le nouveau modèle, le nombre de paramètres à estimer lors de l'apprentissage des modules chute d'un facteur 2 par rapport aux concurrents.

6. Conclusions et perspectives

L'évolution du modèle que nous avons présenté ici évite certaines lacunes de la version précédente : (1) Il n'était pas possible d'exploiter aisément la récursivité des relations de dépendance. Il nécessitait notamment un changement d'architecture (multiplication des modules) selon les extensions que l'on voudrait apporter aux structures syntaxiques qu'il est capable de traiter. Dans la nouvelle version, une fois les relations de dépendance identifiées, le nombre de niveaux de récursivité reste flexible et exerce une influence sur l'apprentissage non sur l'architecture. On peut même espérer obtenir une génération de contours acceptable pour une structure plus profonde que celle du corpus d'apprentissage. (2) La relation entre fonction linguistique et génération de contour est maintenant plus étroite : désormais *un* seul module de génération encode *une* fonction linguistique.

Néanmoins, nous ne disposons que d'une première implémentation du modèle. Mis à part la comparaison numérique des contours prédits et observés et l'analyse de la cohérence des familles de contours, le test crucial devra être une évaluation perceptive de la prosodie générée.

Références

[1] Aubergé, V. *La synthèse de la parole : "des règles aux lexiques"*. PhD thesis, Université Pierre Mendès-France, Grenoble – France, 1991. sous la direction de J. Rouault.

[2] Aubergé, V., Grépillat, T., and Rilliard, A. Can we perceive attitudes before the end of sentences ? The gating paradigm for prosodic contours. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 2, pages 871–874, Rhodes - Greece, 1997.

[3] Bailly, G. Integration of rhythmic and syntactic constraints in a model of generation of French prosody. *Speech Communication*, 8 :137–146, 1989.

[4] Barbosa, P. *Caractérisation et génération automatique de la structuration rythmique du français*. Thèse de troisième cycle, Institut National Polytechnique de Grenoble, Grenoble, France, 1994. Thèse de Doctorat Spécialité Sciences Cognitives sous la direction de Gérard Bailly.

[5] Black, A.W. and Hunt, A.J. Generating f0 contours from tobi labels using linear regression. In *Proceedings of the International Conference on Speech and Language Processing*, pages 1385–1388, 1996.

[6] Cutler, A. and Norris, D. Prosody in situations of communication : salience and segmentation. In *Proceedings of the International Congress of Phonetic Sciences*, volume 1, pages 264–270, Aix-en-Provence, France, 1991.

[7] Emerard, F. Synthèse par diphones et traitement de la prosodie. Thèse de troisième cycle, Université de Grenoble III, Grenoble, France, 1977.

[8] Gee, J.P. and Grosjean, F. Performance structures : a psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15 :411–458, 1983.

[9] Grosjean, F. How long is the sentence ? prediction and prosody in the on-line processing of language. *Linguistica*, 21 :501–529, 1983.

[10] Hirst, D.J. and Di Cristo, A. *Intonation systems : a survey of twenty languages*. Cambridge University Press, Cambridge, 1998.

[11] Holm, B., Bailly, G., and Laborde, C. Performance structures of mathematical formulae. In *Proceedings of the International Congress of Phonetic Sciences*, volume 2, pages 1297–1300, San Francisco, USA, 1999.

[12] Malfrère, F., Dutoit, T., and Mertens, P. Automatic prosody generation using suprasegmental unit selection. In *ESCA/COCOSDA International Workshop on Speech Synthesis*, pages 323–328, 1998.

[13] Möbius, B., Pätzold, M., and Hess, W. Analysis and synthesis of german f0 contours by means of fujisaki's model. *SC*, 13 :53–61, 1993.

[14] Morgan, J.L. and Demuth, K. *Signal to syntax : an overview*. Lawrence Erlbaum Associates, Mahwah, NJ - USA, 1996.

[15] Morlec, Y. *Génération multiparamétrique de la prosodie du français par apprentissage automatique*. PhD thesis, Institut National Polytechnique de Grenoble, Grenoble – France, 1997. Thèse de Doctorat Spécialité Sciences Cognitives sous la direction de Gérard Bailly.

[16] Pasdeloup, V. A prosodic model for french text-to-speech synthesis : A psycholinguistic approach. In Bailly, G., Benoît, C., and Sawallis, T., editors, *Talking Machines : Theories, Models and Designs*, pages 335–348. Esevier B.V., 1992.

[17] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. Tobi : a standard for labeling english prosody. *Proceedings of the International Conference on Speech and Language Processing*, 2 :867–870, 1992.

[18] Traber, C. Fo generation with a database of natural fo patterns and with a neural network. In Bailly, G. and Benoît, C., editors, *Talking Machines : Theories, Models and Designs*, pages 287–304. Elsevier B.V., 1992.