UN MODELE CONNEXIONISTE MODULAIRE POUR L'APPRENTISSAGE DES « GESTES » INTONATIFS

Yann Morlec, Gérard Bailly et Véronique Aubergé

Institut de la Communication Parlée, INPG & Université Stendhal
46, av. Félix Viallet 38031 Grenoble Cedex 01, France
e-mail : (morlec, bailly, auberge)@icp.grenet.fr

1. ABSTRACT

A dynamic model for synthesizing intonation is presented. This model is based on the following assumptions: intonation is the result of superposed and independent prototypical gestures belonging to the diverse linguistic levels: sentence, proposition, syntagm, group, word and phoneme. Prototypical movements are progressively stored in a prosodic lexicon and used by the speaker in given communication tasks. Our current implementation of this model is a modular association of sequential neural networks (SNNs). Each SNN is in charge of the melodic prediction of a specific linguistic level. The resulting melody is the weighted sum of SNNs outputs. We presently focus on the sentence level. We built a corpus of 1932 sentences pronounced with 6 attitudes and various lengths. We then designed SNNs able to perform the expansion of prosodic sentence movement. Preliminary results show that these simple SNNs can give acceptable F0 prediction and keep essential features of each attitude whatever the syllabic length of the sentence.

2. INTRODUCTION

Nous présentons ici une stratégie d'apprentissage d'un modèle dynamique de génération de l'intonation du Français. Notre approche théorique postule que la courbe mélodique est la superposition de « mouvements » intonatifs correspondant aux différents niveaux de représentation du discours (la phrase, la proposition, le syntagme, le groupe, le mot et le phonème). La synthèse de l'intonation résulte de la somme pondérée des prédictions de réseaux connexionistes récurrents modélisant l'expansion des « gestes » prosodiques associés à chacun de ces niveaux. Une attention particulière est portée à la modélisation des contours de phrase véhiculant notamment la position du locuteur vis-à-vis de son discours.

3. CADRE THEORIQUE

Notre analyse de la mélodie est guidée par une approche morphologique. Nous pensons que l'auditeur a des attentes sur ce qu'est une intonation « normale », comme il existe une syntaxe « normale » et qu'il peut ainsi interpréter comme une mise en relief quelque chose qui émerge du fond, qui dévie d'une forme attendue. Cette notion d'attente est à rapprocher aux propositions faites par le mathématicien Petitot-Concordat (Petitot-Concordat, 1986) élève de R. Thom : l'analyse morphologique est un lieu privilégié où notre perception du monde physique peut être guidée - ou biaisée -

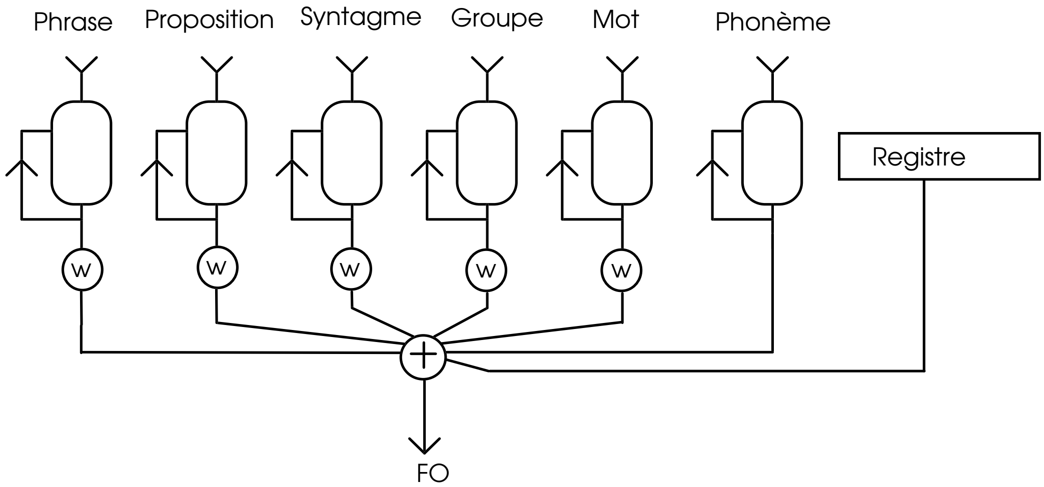

Figure 1: Architecture du modèle de génération des « gestes » intonatifs. Les sorties de chaque module sont exprimées en 1/4 de tons. Le registre donne la fréquence laryngienne moyenne. Les pondérations w permettent d'ajuster l'influence des contours portés.

par la projection de nos représentations mentales préexistantes. C'est par cette boucle action-perception-cognition que nos représentations mentales émergent et s'enrichissent.

Dans cette optique, nous considérons que chaque niveau de description linguistique sur l'axe syntagmatique génère des indices voire des formes qui lui sont propres et qui maintiennent entre elles des contrastes propres à ce niveau. Ainsi la ligne mélodique observée est le résultat de la superposition de ces formes. Nous considérons ici les niveaux de représentation suivants : la *phrase* dont le contour mélodique va véhiculer l'attitude du locuteur. Ce contour de phrase, « geste » porteur, est modulé successivement par les contours de *proposition*, de *syntagme* et de *groupe* qui permettent, en particulier, de rendre compte du degré de cohésion entre les différents constituants de ces niveaux. La micromélodie est l'étape ultime de cette modulation hiérarchique et additive.

L'étude de modèles de superposition pour la génération de l'intonation n'est pas nouvelle. Thorsen propose un modèle ou la ligne mélodique est obtenue par addition de segments de droite (Thorsen, 1980). Fujisaki la décrit comme la superposition de réponses de filtres (Fujisaki, 1971). Notre approche diffère de ces deux exemples par le fait que nous n'imposons pas a priori la forme des contours intonatifs. Par exemple, certains « mouvements » mélodiques de phrase peuvent être complexes et sujets à de brusques variations. De cette description de l'intonation découle le modèle de génération décrit ci-après.

4. UN MODELE D'EXPANSION DU « MOUVEMENT » MELODIQUE

Le modèle dynamique de génération de l'intonation, dans sa version actuelle, consiste en une association modulaire de réseaux connexionistes récurrents chargés de prédire les « gestes » intonatifs des niveaux linguistiques qui leur sont associés. Le contour mélodique final est obtenu par la somme pondérée de ces prédictions (figure 1).

4.1 Des « mouvements » superposés

L'apprentissage du modèle de génération se fait de manière itérative (Aubergé, 1992) - du niveau porteur le plus large vers le niveau porté le plus petit - par la gestion dynamique des pondérations. A la manière des stratégies observées d'apprentissage supervisé de l'intonation (Strömqvist & al., 1995), le modèle va d'abord apprendre à générer un contour mélodique global de phrase. Pour ceci, tous les poids *w* sont initialisés à zéro sauf celui de phrase. Lorsqu'il a acquis une performance suffisante, il va alors moduler ce « geste » mélodique de phrase par l'ajout de la contribution des niveaux inférieurs ; ceci en gelant les paramètres des modules déjà appris et en positionnant le poids *w* correspondant à 1.

La modularité de cette architecture permet donc d'utiliser un corpus approprié à chaque niveau linguistique étudié : le modèle d'expansion du « mouvement » de phrase pour diverses attitudes peut être appris sur un corpus de phrases spécifiques constitué par exemple de mono-mots (cf. paragraphe suivant) alors que, parallèlement, l'apprentissage de la typologie des « mouvements » de groupes peut être effectué sur des phrases assertives à fort branchement syntaxique (Aubergé, 1992).

4.2 Des « mouvements » pondérés

L'existence de pondérations sur les différents niveaux se justifie par les observations suivantes :

- Pour les attitudes correspondant à une reprise mot pour mot de l'assertion de l'interlocuteur (*doute-incrédulité*, *ironie de soupçon),* le locuteur ne ressent pas le besoin de segmenter à nouveau cet énoncé. La réalisation mélodique se restreint alors à un contour de phrase (pondération nulle sur le niveau groupe). Par contre, dans le cas d'une exclamative, la segmentation des contours portés est largement exagérée (pondération supérieure à 1).

- La gestion des poids permet de contrôler dynamiquement l'importance relative des niveaux. Ainsi notre modèle gère facilement les variations de pente de la ligne de déclinaison observées dans des tâches de lecture de textes (Silverman, 1987).

Nos recherches se portent actuellement sur le niveau de la phrase, le découpage de l'énoncé ayant déjà été largement étudié (Morlec & al., 1995).

5. LE NIVEAU PHRASE

Nous nous intéressons ici à la manière par laquelle les locuteurs véhiculent par la voix diverses attitudes. Le but étant de démontrer qu'ils produisent des formes mélodiques bien identifiables.

5.1 Elaboration du corpus

Le corpus *attitudes* comporte 322 phrases de longueur variable (1 à 8 syllabes) prononcées suivant six attitudes intonatives différentes référencées dans (Calbris & al., 1981), soit un total de 1932 phrases. La longueur réduite des phrases du corpus permet de limiter l'influence des contours portés sur la courbe mélodique.

5.2 Caractérisation phonétique

Les réalisations mélodiques sont caractérisées par trois valeurs de F0 par syllabe. Ces valeurs sont prélevées sur la voyelle : on considère qu'en Français, la syllabe ne reçoit qu'un seul mouvement stylisée par une fonction polynomiale du second ordre.

5.3 Description des attitudes

Parmi les six attitudes retenues, deux s'apparentent à une *intonation d'énoncé* où les sentiments personnels du locuteur à l'égard du message qu'il émet sont occultés : l'*affirmation* et la *question simple*. Les quatre autres sont des *intonations de discours* faisant apparaître des informations sur la situation dans laquelle se trouvent les interlocuteurs, sur les dispositions du locuteur vis-à-vis de l'objet dont il parle : l'*exclamation de surprise*, le *doute-incrédulité*, l'*ironie de soupçon* et l'*évidence*. Le *doute-incrédulité* exprime le désaccord partiel et à la limite une négation de ce qui a été préalablement affirmé. Sa mélodie est surtout caractérisée par une mise en relief de l'avant dernière syllabe suivie d'une dernière syllabe montante. Dans l'*ironie de soupçon*, on reprend l'affirmation de l'interlocuteur en lui faisant comprendre qu'on la met en doute, qu'on la contredit. La prosodie de ce type de phrase est caractérisée par un ton haut et, brusquement, sur la dernière syllabe, par une chute de fondamental.

La figure 2 montre la superposition de contours mélodiques (mono-mots essentiellement) pour des phrases de une à huit syllabes correspondant à l'attitude de *doute-incrédulité*. Ces courbes, comme celles associées aux autres attitudes étudiées, donnent les caractéristiques essentielles de l'expansion du « mouvement » prosodique en fonction du nombre de syllabes. Ces contours paraissent se diviser en deux parties : une phase finale de capture très prototypique sur une à deux syllabes précédée d'une phase d'amorce caractérisée par une valeur initiale assez stable et suivie d'une relaxation.

Cette analyse met en lumière l'existence de contours mélodiques globaux qui ne sont pas « ancrés » sur les niveaux inférieurs de la phrase. Pour les intonations d'énoncé, on peut néanmoins remarquer l'apparition de modulations supplémentaires lorsque la phrase dépasse typiquement 5 à 6 syllabes. Cette modulation est due, dans notre modèle, à la nécessité d'apporter par l'adjonction de contours portés des informations sur découpage de l'énoncé en groupes.

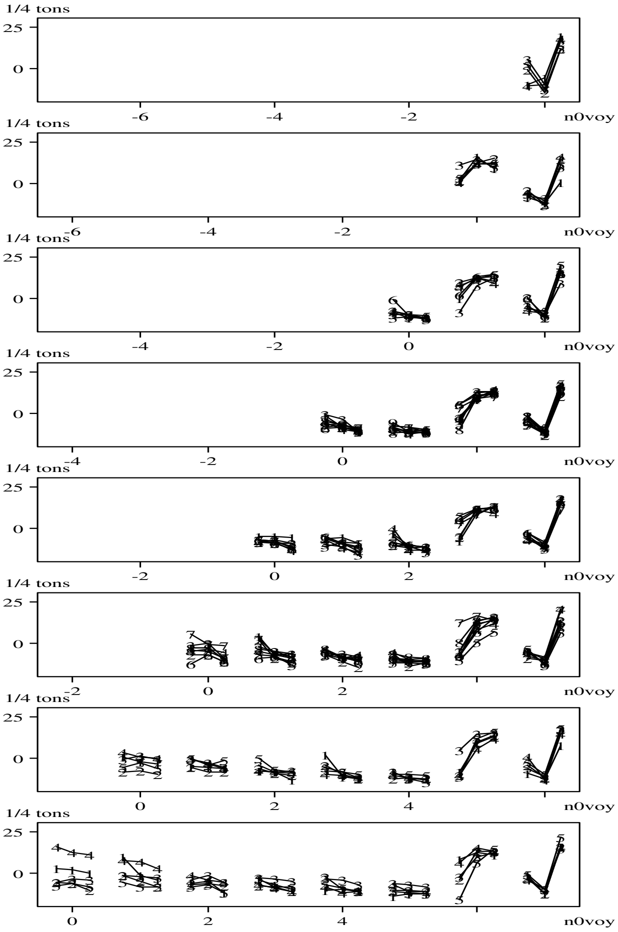

Figure 2 : Contours mélodiques pour l'attitude de doute-incrédulité appliquée à des phrases de 1 à 8 syllabes.

5.4 Architecture du module *phrase*

Il comporte un ensemble de réseaux de neurones de type Elman (Elman, 1988) destinés à prédire les contours mélodiques véhiculant les six attitudes enregistrées dans le corpus *attitudes.*Ces réseaux dynamiques ont une architecture semblable comportant :

- En entrée, deux rampes linéaires décroissantes relative (de 1 à 0) et absolue (du nombre de syllabes à 0). Elles permettent de gérer la séquentialité du « mouvement » et de déclencher la phase de capture au bon moment.
- Deux couches cachées interconnectées dans les deux sens.
- La sortie correspond à la F0 stylisée.

5.5 Apprentissage

Les corpus d'apprentissage de ces réseaux sont constitués de l'ensemble des phrases mono-mot de 1 à 6 syllabes, soit 18 phrases par attitude (les mono-mots permettent de neutraliser l'incidence des contours portés). Les phrases de 7 et 8 syllabes sont, quant à elles, destinées à vérifier leur capacité de généralisation (corpus de test).

6. RESULTATS

Sur le corpus d'apprentissage, les prédictions générées par le module *phrase* sont proches des moyennes des échantillons d'apprentissage (comprises entre μ-σ et μ+σ).

Pour l'attitude de *doute-incrédulité*, les figures ci-après permettent de comparer les prototypes prédits par le réseau (figure 4) avec la moyenne des contours d'apprentissage (figure 3). Notez que le réseau a su reproduire fidèlement les caractéristiques essentielles de ce type d'attitude : on retrouve la mise en relief de l'avant-dernière syllabe et le « mouvement » montant sur la dernière.

La généralisation des réseaux à été étudiée sur les phrases du corpus de test de 7 et 8 syllabes (figure 4). Les résultats montrent que la phase de capture est toujours bien mémorisée. L'expansion du « mouvement » se réalise correctement sur l'étape de préparation.

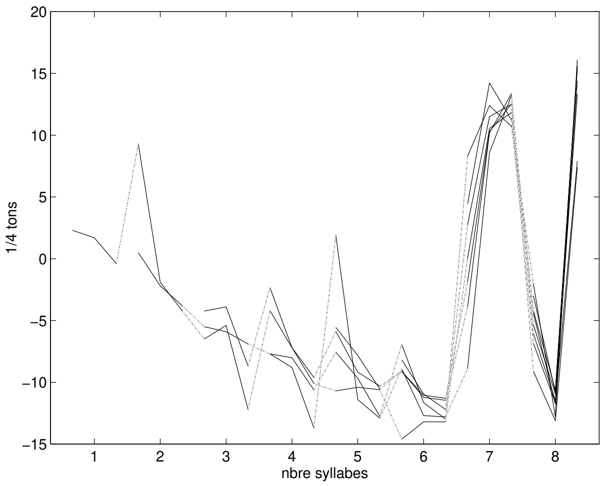

Figure 3 : Moyennes des F0 stylisées de phrases monomot de 1 à 8 syllabes pour le doute-incrédulité.

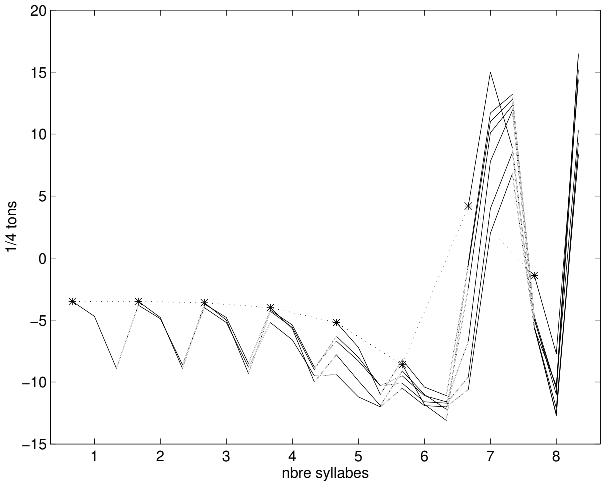

Figure 4 : Prédictions mélodiques du réseau (1 à 6 syllabes) et généralisation à 7 et 8 syllabes.

7. CONCLUSIONS

L'objectif de notre étude est double :

(a) démontrer que notre modèle morphologique de la prosodie peut émerger d'un apprentissage hiérarchique de contours. Cette focalisation de l'apprentissage sur divers niveaux linguistiques est à rapprocher de la théorie Spotlight (Strömqvist & al., 1995) : la prosodie sert à enrichir la compétence linguistique de l'enfant par « zooms » successifs.

(b) démontrer que cette approche produit des systèmes de génération de la prosodie performants et évolutifs.

Ce modèle de superposition de « gestes » prosodiques est semblable à des modèles de contrôle gestuels de la parole (la mâchoire comme organisateur syllabique est porteuse des gestes labiaux et linguaux) et des modèles d'écriture manuscrite (le bras gère le déplacement linéaire et porte les mouvements de la main formant les lettres).

8. BIBLIOGRAPHIE

Petitot-Concordat J. (1986) *Le « morphological turn » de la phénoménologie. Chapitre I, II et III de morphogénèse du sens II*, Centre d'analyse et de Mathématique Sociales, EHESS - CNRS

Thorsen N. (1980) A study of the perception of sentence intonation - evidence from Danish*, Journal of the Acoustical Society of America*, 67(3):1014-1030

Fujisaki H. and Sudo H. (1971) A generative model for the prosody of connected speech in japanese*, Annual Report of Engineering Research Institute,* 30:75-80

Aubergé V. (1992) Developing a structured lexicon for synthesis of prosody, In Bailly, G., Benoît, C., editors, *Talking Machines: Theories, Models and Design*, Elsevier B.V. 307-321

Strömqvist S., Peters A. & Ragnarsdóttir H. (1995) Particles and prepositions in Scandinavian child language development : effects of prosodic spotlight ? *International Congress of Phonetic Sciences*, Stockholm, Sweden, 4:38-45

Silverman K. (1987) *The Structure and Processing of Fundamental Frequency Contours,* Ph.D. thesis, Cambridge University, Cambridge UK

Morlec Y., Aubergé V., & Bailly G. (1995) Evaluation of automatic generation of prosody with a superposition model, *International Congress of Phonetic Sciences*, Stockholm - Sweden, 4: 224-227

Calbris G. & Montredon J. (1981) *Oh là là. Expression intonative et mimique. Livre du professeur*, CLE international

Elman J.L. (1988) *Finding structure in time*, CRL Tech. Rep. 8801, University of California, San Diego