

Génération automatique de la prosodie pour la synthèse à partir du texte : le système KALI

Anne Lacheret-Dujour, Michel Morel

Laboratoire Crisco - Université de Caen - Esplanade de la Paix, 14032, Caen cedex

Tél : ++33 (0)2 31 56 56 27 – Fax : ++33 (0)2 31 56 54 27

anne.lacheret@crisco.unicaen.fr ; michel.morel@crisco.unicaen.fr

ABSTRACT

Kali, a French-speaking text-to-speech synthesis software package created for visually handicapped people, is the result of a collaboration between University and the private sector. The input text goes through a succession of 5 modules (preprocessing, syntactic analysis, prosodic generation, phonemisation, acoustico-phonetic processing) and is then pronounced. Its best feature is intelligibility at rapid delivery. In this paper, syntactic and prosodic modules are presented.

1. INTRODUCTION

Si les structures syntaxique et prosodique ne sont pas nécessairement congruentes, il n'en reste pas moins qu'elles demeurent associées, dans la mesure, bien sûr, où l'intonation respecte par ailleurs un jeu de contraintes textuelles et rythmiques fondamentales. La segmentation en constituants syntaxiques (ici dénommés "tronçons") représente donc un point d'ancrage précieux pour poser une structure prosodique de base en vue de la génération automatique de la prosodie. Nous reprenons pour notre analyse les principes formulés dans le cadre des grammaires de dépendance¹, selon lesquels la phrase syntaxique peut être vue comme un processus de mise en relation mémorielle, modélisable et généralisable dans une perspective de traitement automatique.

La génération automatique de la prosodie suppose ensuite deux niveaux de représentation : (i) le niveau qualitatif, fondé sur une modélisation phonologique de l'intonation, consiste à générer un jeu de marqueurs abstraits pour rendre compte d'une structure intono-syntaxique hiérarchisée, (ii) le niveau quantitatif a pour fonction de faire correspondre à ces marqueurs les corrélats acoustico-phonétiques pertinents : aux différents types de prééminences associées à cette hiérarchie correspondent des variations plus ou moins fortes de fréquence fondamentale, d'intensité et de durée².

¹ Voir [Bai86] pour une des premières applications à la synthèse du français.

² Pour une présentation détaillée du traitement, voir [Mor01].

2. ANALYSE MORPHOSYNTAXIQUE

Notre analyse s'effectue en trois passes, la segmentation en phrases et en mots, l'étiquetage de ceux-ci et leur regroupement en tronçons syntaxiques, eux-mêmes mis en relation les uns avec les autres, la dernière étape se fondant sur la modélisation de processus de propagation et de déductions contextuelles.

2.1 Segmentation : du paragraphe au mot

En vue d'un traitement prosodique textuel, les unités à isoler sont respectivement les paragraphes, les phrases à l'intérieur des paragraphes et les mots dans les phrases. Si les deux premiers segments sont simples à identifier – un saut de ligne indique un nouveau paragraphe, la phrase est repérée par un signe de ponctuation terminale³ – il n'en va pas de même pour le dernier [Fuc93]. Les critères typographiques d'identification sont, en effet, nécessaires mais non suffisants (*l'homme vs. aujourd'hui, porte-monnaie vs. voulez-vous*). Compte tenu de cette difficulté, notre principe de segmentation est le suivant : les caractères de ponctuations, apostrophes, parenthèses, guillemets, etc. sont considérés comme des séparateurs de mots, les exceptions étant codées dans un dictionnaire.

2.2 Etiquetage des mots

Pour ce traitement, outre deux bases lexicales recensant les exceptions aux règles générales (190 formes figées et 100 homographes-hétérophones), les dictionnaires codant les mots grammaticaux et les verbes constituent le pilier de l'analyse. En effet, peu nombreux et très stables, ils organisent pour une large part les relations entre les constituants syntaxiques. Les verbes, notamment, constituent la base de notre segmentation en tronçons : toute séquence qui ne se construit pas autour d'une base verbale est considérée par défaut comme nominale.

Dans tous les cas, les entrées polycatégorielles font l'objet de plusieurs étiquetages (ex : *a priori* adverbe

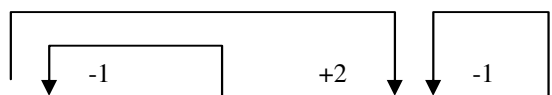
³ Sauf lors de l'identification d'un sigle (en contexte où plusieurs points encadrent un groupe de lettres contiguës).

ou nom, *président* : nom masculin singulier ou verbe pluriel, 3^{ème} personne).

60 règles de déduction contextuelle, ensuite, permettent d'étiqueter les mots non encore catégorisés, ou faisant l'objet d'un étiquetage multiple, en utilisant les informations fournies par le contexte et en les propageant par déduction. Ainsi, une entrée étiquetée [déterminant] ou [pronom objet], est recatégorisée [pronom] lorsqu'elle est précédée d'un pronom sujet (*il l'entraîna dehors*). Enfin, un suffixe étant, dans bien des cas, porteur d'une information morphosyntaxique [Bou97], une base de 600 suffixes est consultée, le cas échéant, pour résoudre les ambiguïtés qui demeurent à l'issue de ces différents traitements.

2.2 Segmentation des tronçons et mise en relation

En premier lieu, quelques déductions simples, effectuées directement par le programme, conduisent à regrouper les mots fortement liés syntaxiquement en tronçons (ex : [dét+n+adj], [aux+v], etc. Soit la segmentation suivante : (*l'a.d.n.*) (*des hommes célèbres*) (*intéresse*) (*le président*). La mise en relation des tronçons entre eux ensuite, clé de voûte à la hiérarchisation prosodique, constitue une opération beaucoup plus complexe qui fait appel à un jeu de 41 règles. En pratique, l'analyseur dispose de différentes mémoires – ou piles – chacune dédiée à une relation syntaxique particulière (ex : relation SV, N-CN, etc.). La chaîne est analysée de gauche à droite, chaque nouveau tronçon, inséré dans une mémoire, est empilé et devient le premier candidat pour une mise en relation éventuelle. Un groupe qui n'est plus en attente d'un autre groupe, parce que déjà relié, est effacé de la mémoire⁴. Soit pour la phrase ci-dessus, le traitement suivant : (*l'a.d.n.*) est mémorisé comme sujet possible en attente d'un tronçon verbal, (*des hommes célèbres*) est relié au premier tronçon (relation N-CN) et retiré de la pile ; (*intéresse*) valide (*l'a.d.n.*) comme sujet, celui-ci est donc effacé de la mémoire. Enfin, pour l'homographe hétérophone (*le président*), la règle appliquée est la suivante : [nom/verbe] en contexte gauche verbal est réécrit nominal et relié au tronçon qui le précède immédiatement.



(*l'a.d.n.*)(*des hommes célèbres*)(*intéresse*)(*le président*)

Figure 1 : Exemple de mise en relation syntaxique

⁴ Ce concept d'attente est à rapprocher de celui de saturation valancielles formulé dans [Tes59].

3. CALCUL DE LA PROSODIE

L'ensemble du traitement prosodique (qualitatif et quantitatif) est effectué ici par un jeu de 90 règles, construites sur les bases de l'observation acoustique de deux corpus de lecture oralisée⁵. Le choix des corpus – textes et non phrases isolées – a été décisif pour vérifier l'hypothèse selon laquelle la dimension textuelle du message à synthétiser représente un paramètre essentiel dans la construction de la structure prosodique⁶. Cette dernière y est présentée comme le produit de plusieurs composants imbriqués et de portée variable : un composant global se manifeste sur l'ensemble de l'énoncé et sur les groupes qui le constituent, un composant local s'exprime par la prééminence de syllabes accentuées de différente nature.

3.1 Analyse qualitative

De façon classique, les tronçons syntaxiques sont considérés comme des groupes accentuels virtuels (GA)⁷ qui servent de base pour générer une structure intonative hiérarchisée. Les contraintes appliquées ensuite se résument à trois principes : (un principe de cohésion textuelle (PCT), un principe d'alignement syntaxique (PAS) et un principe de bonne formation rythmique (PBR).

Suivre le premier, nous amène à manipuler trois unités de traitement : le paragraphe, la phrase et le groupe de souffle⁸. Ces 3 segments se caractérisent par une déclinaison et une pause terminale de durée variable (tableau 1).

Tableau 1 : Frontières prosodiques et organisation textuelle (le degré de frontière est indiqué de façon croissante).

Niveau 1 : FTPg Frontière terminale de §
Niveau 2 : FTP_h Frontière terminale de phrase
Niveau 3 : FCGS Frontière continuative de groupe de souffle

L'application de PAS tient compte ensuite de deux types de dépendance syntaxique (contiguë, DC ou à distance, DD). Dans les deux contextes, pour deux tronçons linéairement adjacents, le premier est ponctué par une prééminence accentuelle terminale associée à un allongement. Cet allongement est plus marqué dans les DD, pour lesquelles une pause est, en outre, insérée, sa durée étant proportionnelle au nombre de syllabes à

⁵ Extrait d'un roman policier (520 mots), article de presse (500 mots), un locuteur, deux lectures pour un même texte.

⁶ Pour des approches comparables, voir [Slu93], [Mert01].

⁷ Chaîne de syllabes dont la dernière est marquée par un accent terminal de mots.

⁸ Unité marquée à sa droite par un signe de ponctuation interne à la phrase.

parcourir dans la phrase pour relier l'unité régie à son régissant. Deux nouvelles frontières sont ainsi définies :

Tableau 2. Frontières prosodiques et relations de dépendance syntaxique

Niveau 4 : FCGI Frontière Continuitive Majeure de groupe Intonatif (relation de dépendance à distance)
Niveau 5 : FcGI Frontière Continuitive Mineure de Groupe intonatif (relation de contiguïté)

La prise en compte de PBR, enfin, consiste à insérer ou effacer certains marqueurs accentuels ou pausals en appliquant les contraintes de régulation temporelle (CRP) et accentuelle (CRA)⁹. CRP conduit à l'effacement d'une pause générée entre deux constituants non reliés si moins de 8 syllabes séparent les groupes qui contractent cette relation, l'allongement syllabique étant préservé (*ex : (il promène) (ses enfants) (dans le jardin) vs. (il promène) (les enfants) (de Nathalie) (et Vincent) # (dans le jardin)*). Dans le deuxième exemple, la pause est maintenue, de même nature que la pause générée par une virgule. Nous regroupons ainsi dans une catégorie unique (FCGS) tous les groupes de souffle de taille inférieure à la phrase. Soit, pour exemple, les règles ci-dessous :

- ((tronçon) et (virgule)) → FCGS
- (tronçon)+((tronçon) et (distance ≥ 8)) → FCGS
- (tronçon)+((tronçon) et (0 < distance < 8)) → FCGI
- (tronçon)+((tronçon) et (distance = 0¹⁰)) → FcGI

Selon CRA, un groupe constitué d'un nombre de syllabes trop important (4 syllabes ou plus) est accentué sur la première syllabe à attaque consonantique de son premier mot lexical (*une activité valorisante*), c'est-à-dire donnant lieu à la formation d'un pied métrique¹¹ et donc à la définition d'une sixième frontière (**Niveau 6 : FPM**). Le tableau 3 résume notre hiérarchie intonative.

Tableau 3 : Hiérarchie des frontières intonatives et paramètres phonétiques associés

Niveau	Décl	Pause	Allgt	Proém (F0, I)
1	FTPg	FTPg	FTPg	
2	FTPPh	FTPPh	FTPPh	
3	FCGS	FCGS	FCGS	FCGS
4			FCGI	FCGI
5			FcGI	FcGI
6				FPM

3.1 Analyse quantitative

L'interprétation phonétique des objets posés par l'analyse qualitative consiste d'abord à calculer une déclinaison pour chacun des 3 niveaux indiqués dans le tableau 1. La mélodie de l'énoncé est ensuite synthétisée en superposant la partition intonative de chaque niveau. La réinitialisation subséquente n'est pas calculée mais dérive du calcul des pentes. Ces dernières diminuent en valeur absolue dès que la taille des groupes dépasse 5 syllabes afin que l'amplitude maximale de variation ne dépasse jamais le registre d'un locuteur humain en situation de lecture. Le paramètre de pente P dépend donc du nombre de syllabes S et se décompose en deux paramètres : la pente maximale P et l'amplitude A , d'où le choix d'une fonction homographique respectant les conditions aux limites :

$$p = \frac{A}{s + \frac{A}{P}}$$

- Si $s \rightarrow 0$, alors $p \rightarrow P$ (pente maximale)
- Si $s \rightarrow \infty$, alors $p \cdot s \rightarrow A$ (amplitude max de variation)

Le calcul acoustico-phonétique des proéminences locales ensuite, distinguant l'accentuation non terminale (accent secondaire) de l'accentuation finale (accent primaire) recense 4 paradigmes :

- Syllabes inaccentuées des mots lexicaux (IML) ;
- Syllabes inaccentuées des mots grammaticaux (IMG)¹² ;
- Syllabes accentuées démarcatives de pieds métriques (FPM) ;
- Syllabes accentuées finales de groupes.

Le tableau 4 résume les principaux paramètres acoustiques associés aux différents types de frontières et de paradigmes syllabiques.

⁹ Voir [Pas90], [Lac99] pour une présentation détaillée des contraintes.

¹⁰ La distance est nulle quand les tronçons en relation sont linéairement adjacents (paramètres de relation +1 ou -1).

¹¹ Unité bornée à droite par un accent rythmique non terminal de groupe syntaxique.

¹² Fréquemment réduites y compris en situation de lecture.

Tableau 4 : Correspondance acoustique (unités logarithmiques) de la hiérarchie intonative, où la dernière syllabe de la phrase possède 2 valeurs correspondant à un glissando vocalique.

Niveau	F0	Intens.	durée
IML	0	0	0
IMG	-6	-6	-8
FPM	12	8	0
FcGI	18	10	18
FCGI	24	10	24
FCGS	32	0	32
FTP ¹³	-28 -56	-18 -36	20 40

A la fin de l'analyse quantitative, il reste à mettre en œuvre le modèle en positionnant sur chaque noyau syllabique identifié les marqueurs nécessaires au module acoustico-phonétique qui se chargera de les interpoler plus finement, phonème par phonème.

3. BILAN ET PERSPECTIVES

Nous nous sommes attachés à décrire ici la construction d'un modèle de congruence syntaxe-prosodie pour la génération automatique de la parole. Pour le volet syntaxique, nous avons mis en œuvre un traitement qui repose sur la mise en relation syntaxique de segments, *i.e.* sur la mise en évidence de processus dynamiques en nombre fini à l'origine de la production ou de l'interprétation des structures. Dans cette approche, la phrase est considérée comme le codage linéaire d'une représentation dépendantielle abstraite, qui doit obéir à une contrainte cognitive fondamentale, de minimisation de l'effort mémoriel dont la prosodie porte les traces. Selon cette dernière, la distance entre deux tronçons syntaxiquement contigus doit être minimale dans l'ordre linéaire (voir la contrainte du lien minimal en Grammaire Générative). Pour le traitement prosodique ensuite, notre analyse constitue la première phase d'un travail à long terme sur l'étude et la modélisation des contraintes textuelles dans la construction de la structure intonative. Pour l'heure, nous avons pu faire émerger une hiérarchie dans les paramètres mobilisés (durée et déclinaison) pour la production d'unités distinctes comme la phrase et le paragraphe. Du point de vue de l'agrément d'écoute, qui constitue l'un de nos objectifs prioritaires étant donné le public concerné par Kali, cette approche permet de limiter la monotonie de la voix synthétique associée à la concaténation de patrons prosodiques de phrases réitérés sur l'ensemble d'un texte.

Puisqu'il s'agit de travaux qui se situent dans le cadre de la recherche appliquée, où les critères d'efficacité, de fiabilité, d'évolutivité et de maintenance sont décisifs, nous ne saurions conclure cette communication

¹³ En modalité déclarative.

sans quelques mots sur l'évaluation. Différents outils ont été développés au laboratoire pour optimiser l'écriture, l'évaluation et la correction des règles (corpus de tests et traçage automatique des règles par exemple). Une évaluation "maison" a pu ainsi être menée, mettant au jour un nombre relativement important d'erreurs – en cours de correction – que l'on peut regrouper en 2 familles : (i) mauvais regroupements syntaxiques – ex : *je dis ce que je pense* (locution *ce que*), *il parle de vous y laisser* (déduction locale *vous y*), *ils avaient désormais un semblant de stratégie* (polycatégorie *semblant*), (ii) mauvaise détection de pauses – ex : *il traverse La Rochelle d'un bout # à l'autre* (expression figée *d'un bout à l'autre*). Par leurs répercussions sur les liaisons, les pauses et les accents démarcatifs, ces erreurs nuisent à la compréhension et à l'agrément d'écoute. Poursuivre cette évaluation constitue l'un de nos objectifs prioritaires de recherche à moyen terme en classant les erreurs par type et par ordre de priorité, afin de les traiter ultérieurement, ceci en respectant une contrainte évidente : les modifications apportées doivent produire les résultats escomptés, sans pour autant générer des erreurs imprévues.

BIBLIOGRAPHIE

- [Bai86] Bailly G. (1986), "Un modèle de congruence relationnel pour la synthèse de la prosodie du français", *15èmes Journées d'étude sur la parole*, Aix-en-Provence, 75-78.
- [Bou97] Boula de Mareüil P. (1997), *Etude linguistique appliquée à la synthèse de la parole à partir du texte*, Thèse de Doctorat, Paris XI.
- [Fuc93] Fuchs C., Danlos L., Lacheret-Dujour A., Luzzati D., Victorri B. (1993), *Linguistique et traitement automatique des langues*, Paris, Hachette.
- [Lac99] Lacheret-Dujour A., Beaugendre B. (1999), *La prosodie du français*, Paris, Editions du CNRS.
- [Mer01] Mertens P., Auchlin A., Goldman J.P., Grobet A. (2001), "L'intonation du discours : une implémentation par balises ; motifs et premiers résultats", ce vol.
- [Mor01] Morel M., Lacheret-Dujour A. (2001), "Le logiciel de synthèse vocale Kali : de la conception à la mise en œuvre", à paraître dans *TAL*, Ch. D'Alessandro (éd.), Paris, Hermes.
- [Pas90] Padeloup V. (1990), *Modèle de règles rythmiques du français appliqué à la synthèse de la parole*, Thèse de Doctorat, Aix-en-Provence.
- [Slu93] Sluijter A., Terken J.M.B. (1993), "Beyond sentence prosody : paragraph intonation in Dutch", *Phonetica* 50, 180-188.
- [Tes59] Tesnière L. (1959), *Éléments de syntaxe structurale*, Paris, Klincksieck.

