Génération automatique de la prosodie dans le système de synthèse vocale Kali : de la modélisation phonologique à l'implémentation des paramètres acoustiques

Anne Lacheret-Dujour, Michel Morel

CRISCO
Université de Caen, 14032 Caen cedex
Tél.: ++33 (0)231565627 - Fax : ++33 (0)231565427
Mél : anne.lacheret@crisco.unicaen.fr ; michel.morel@crisco.unicaen.fr – http//www.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, prosodic processing is presented from the phonological representation of intonation to the acoustic processing.

1. INTRODUCTION

Le module de génération automatique de la prosodie présenté ici repose sur une modélisation phonologique de la structure intonative qui sert d'entrée au traitement phonétique des données et à l'implémentation des paramètres acoustiques.

La représentation phonologique de l'intonation repose sur trois traitements : textuel, intonosyntaxique et rythmique. La prise en compte de contraintes textuelles, ressentie comme nécessaire aujourd'hui [Slu93] [Mer01], a pour objet de limiter l'inévitable monotonie de patrons intonatifs de phrases réitérés de manière invariante sur l'ensemble du texte à synthétiser. Sous l'angle intonosyntaxique, les phrases ont fait l'objet d'une segmentation préalable en tronçons (ou chunks[1]) mis en relation les uns avec les autres par des calculs de dépendance [Ver99] [Lac01] [Mor01][2]. Ces groupes syntaxiques constituent les primitives d'entrée à la dérivation prosodique : les chunks sont considérés comme des groupes accentuels potentiels. Les frontières prosodiques dérivent du calcul des relations de dépendance syntaxique entre les chunks ainsi constitués et font l'objet de réajustements rythmiques simples et limités. Ces différents traitements conduisent à poser un jeu de marqueurs abstraits rendant compte d'une structure intonative hiérarchisée.

Les corrélats acoustico-phonétiques des marqueurs associés à la structure intonative sont ensuite calculés pour générer la prosodie de la parole synthétique.

L'ensemble du traitement (phonologique et acoustique) est effectué par un jeu de 90 règles, construites sur les bases de l'observation acoustique[3] de deux corpus de lecture oralisée [Van99], l'extrait d'un roman policier (520 mots) et un article de presse (500 mots)[4]. Dans l'approche retenue, la structure prosodique est présentée comme le produit de plusieurs composants imbriqués de portée variable : un composant global se manifeste sur l'ensemble de l'énoncé et sur les groupes qui le constituent, un composant local est associé à la proéminence des syllabes accentuées démarcatives (accent primaire) et, le cas échéant, internes de mot (accent secondaire).

Nous présentons ainsi une approche superpositionnelle de l'intonation[5] dans laquelle les accents sont considérés comme des proéminences locales subordonnées à l'intonation de groupes de différentes natures (intonation de paragraphe, de phrase, de groupe de souffle), elle-même modélisée par une ligne de déclinaison.

2. LA REPRÉSENTATION PHONOLOGIQUE

Hormis le pronom sujet, toujours atone, les constituants dérivés de l'analyse syntaxique automatique sont considérés comme des unités virtuellement accentuables sur leur dernière syllabe pleine :

(p1) *Marie viendra ce soir*
(p2) *Elle viendra ce soir*

A partir de cette représentation accentuelle de base, il s'agit de dériver une structure intonative hiérarchisée et rythmiquement bien formée. Dire que la structure accentuelle s'articule autour de trois types de contraintes fondamentales, textuelles, syntaxiques et rythmiques, amène

[1] Voir [Abn92] pour l'introduction du concept.
[2] Pour ce type d'approche, voir aussi [Bou97].
[3] Utilisation du logiciel Momel [Hir93].
[4] Le choix du corpus (texte et non phrases isolées) a été fondamental pour vérifier l'hypothèse selon laquelle la dimension textuelle du message à synthétiser représente un paramètre essentiel dans la construction prosodique.
[5] Voir [Lac99] et [Ros99] pour une présentation des différents types de modèles.

à manipuler six classes de frontières intonatives (*cf. infra*, § 2) tenant compte de ces différents niveaux d'organisation.

2.1. Contraintes textuelles

La prise en compte des contraintes textuelles repose sur la différenciation de trois unités de traitement : le paragraphe, la phrase typographique et le groupe de souffle. Le dernier, déterminé par la ponctuation interne des phrases, est démarqué à droite par une virgule, un point-virgule ou deux points. Ces trois unités se caractérisent par une ligne de déclinaison et une pause terminale de durée variable – la pause la plus forte est attribuée à l'unité 'paragraphe'. D'où un premier jeu de frontières, hiérarchisées comme suit :

*Niveau **1** : **FTPg***	*Frontière Terminale de Paragraphe*
*Niveau **2** : **FTPh***	*Frontière Terminale de Phrase*
*Niveau **3** : **FCGS***	*Frontière Continuative de Groupe de Souffle*

2.2. Contraintes d'alignement syntaxique

Les contraintes d'alignement syntaxique dérivent du calcul des dépendances syntaxiques – contiguës (p3) ou à distance (p4) :

(p3) *(Martin)$_a$ (ne viendra pas)$_b$*

(p4) *(Martin)$_a$ (comme tu le sais sans doute)$_b$ (ne viendra pas)$_c$*

Dans le contexte (p3), la relation de contiguïté syntaxique entre un chunk 'a' et un chunk 'b' linéairement adjacents est marquée par une proéminence accentuelle terminale associée à un allongement syllabique. Dans le contexte (p4), l'élément régi 'a' et son recteur 'c' n'entrent pas dans une relation linéaire de contiguïté. Cette construction est marquée par une proéminence accentuelle associée à un allongement plus prolongé de la dernière syllabe du constituant 'a' et par l'insertion d'une pause dont la durée est proportionnelle au nombre de syllabes à parcourir dans la phrase pour relier les unités entretenant un rapport de rection – ici : une distance de 6 syllabes. Deux nouvelles frontières sont ainsi produites :

*Niveau **4** : **FCGI***	*Frontière Continuative Majeure (dépendance à distance)*
*Niveau **5** : **FcGI***	*Frontière Continuative Mineure (relation de contiguïté)*

2.3. Contraintes rythmiques

Les contraintes rythmiques, qui amènent à insérer ou effacer certains marqueurs accentuels et pausals, résultent de l'application de deux principes : un principe de régulation temporelle et un principe de régulation accentuelle. Suivre le premier consiste à effacer une pause générée entre deux constituants non reliés lorsque moins de 8 syllabes les séparent :

(p5) *(je suis né) (à Alger) (en 1943)*

(p6) *(je suis né) (lors de l'insurrection algérienne) # (en 1943)*

Dans (p5), si l'allongement de la dernière syllabe du deuxième chunk reste important, la pause est en revanche effacée. Selon le second principe [Pas90], un groupe constitué d'un nombre de syllabes inaccentuées trop important (≥ 4) est accentué sur la première syllabe à attaque consonantique de son premier mot lexical *(d'interminables escalators)*. Ce qui nous amène à poser la dernière frontière :

*Niveau **6** : **FPM***	*Frontière de pied métrique*

Table 1 : Hiérarchie des frontières et marqueurs acoustiques associés

Niveau hiérarchique	Déclinaison	Pause	Allgt	Proém. (f0, I)[6]
1	FTPg	FTPg	FTPg	
2	FTPh	FTPh	FTPh	
3	FCGS	FCGS	FCGS	FCGS
4			FCGI	FCGI
5			FcGI	FcGI
6				FPM

où :

- Les pauses résultent de contraintes typographiques et syntaxiques.
- Une ligne de déclinaison est associée aux groupes intonatifs ponctués par des pauses.
- L'allongement caractérise toutes les frontières terminales.
- Les proéminences accentuelles dérivent de principes syntactico-rythmiques.

3. LE TRAITEMENT ACOUSTIQUE

La génération des paramètres acoustiques suppose un choix préalable d'unités de mesure aussi pertinentes que possible. Nous avons choisi pour la f0, l'intensité et l'allongement, des unités logarithmiques de granularité suffisamment fine pour assurer une bonne précision (table 2). Afin de donner une idée de cette granularité, nous fournissons un équivalent tonal, par extension de la notion de ton utilisée en musique (par exemple, 1/8^e de ton en intensité correspond à un rapport identique à 1/8^e de ton en hauteur). Les pentes sont exprimées dans les mêmes unités logarithmiques relativement au nombre de syllabes du groupe concerné et les pauses en nombre de phonèmes[7].

[6] Fréquence fondamentale (f0) et intensité (I).

[7] Un phonème dure environ 100ms à vitesse de phonation modérée.

Table 2 : Unités utilisées pour les paramètres acoustiques

Paramètres acoustiques	Rapport correspondant à une unité	Equivalent tonal
F0	1,00726	1/16e ton
I	1,0146	1/8e ton
Allgt	1,0146	1/8e ton

Les variations logarithmiques des trois types de paramètres acoustiques sont calculées par rapport aux syllabes inaccentuées de référence contenues dans la base de diphones préenregistrée. Un jeu de paramètres phonétiques est défini et quantifié pour chacun des niveaux de la hiérarchie intonative, variable en fonction des modalités de phrase (déclarative, interrogative, suspensive, exclamative).

3.1. Le calcul de la ligne de déclinaison

L'interprétation phonétique du modèle phonologique consiste d'abord à calculer une déclinaison pour chacun des 3 niveaux indiqués en *supra*, § 2.1. La mélodie de l'énoncé est ensuite synthétisée en superposant la partition intonative de chaque niveau. La réinitialisation subséquente dérive directement du calcul des pentes qui diminuent en valeur absolue dès que la taille des groupes dépasse 5 syllabes. Etant donné cette approche, 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 maximale *A*, d'où le choix d'une fonction homographique respectant les conditions aux limites :

(f1)
$$p = \frac{A}{s + \frac{A}{P}}$$

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

Table 3 : Paramètres de déclinaison (modalité déclarative)

Niveau hiérarchique	Amplitude maximale	Pente maximale
1. FTPg	-3	-0,75
2. FTPh	-6	-1,5
3. FCGS	-12	-3

Par exemple, dans notre approche superpositionnelle, une phrase isolée formée d'un seul groupe de souffle prend pour paramètres respectifs la somme des paramètres des 3 niveaux : -21 et -5,25. Si elle comporte 5 syllabes, d'après la fonction (f1), p = -2,33 unités de hauteur par syllabe.

3.2. Le modèle de pauses

Pour la modalité déclarative et les pauses typographiques ou syntaxiques, nos modèles de pauses se déclinent comme suit :

Table 4 : Paramètres des pauses (modalité déclarative)

Niveau hiérarchique	Paramètres de durée	Correspondance en millisecondes
1. FTPg	4	400
2. FTPh	6	600
3. FCGS	distance[8] × 0,25 (maximum 6)	200 à 600

Les pauses de niveau 1 et 2 s'ajoutent ; autrement dit la dernière phrase d'un paragraphe est suivie d'une pause de 10 unités (soit une seconde à vitesse de phonation moyenne) contre 6 unités pour une phrase située à l'intérieur d'un paragraphe. Evidemment, la pause de niveau 3 ne peut pas coexister avec les deux précédentes puisqu'elle correspond à une frontière non terminale.

3.3. Le traitement des proéminences locales

Quatre paradigmes syllabiques sont distingués pour effectuer le calcul acoustico-phonétique des proéminences locales :

- Syllabes inaccentuées des mots lexicaux (IML) ;
- Syllabes inaccentuées des mots grammaticaux (IMG)[9] ;
- Syllabes accentuées démarcatives de pieds métriques (FPM) ;
- Syllabes accentuées finales de groupes.

A l'issue de ce traitement, nos principaux paramètres acoustiques sont les suivants pour la modalité déclarative :

Table 5 : Correspondance acoustique (unités logarithmiques) de la hiérarchie intonative, où la dernière syllabe de la phrase possède deux valeurs correspondant à un glissando vocalique (modalité déclarative).

Niveau	**F0**	**Intens.**	**durée**
IML	0	0	0
IMG	-6	-6	-8
FPM	12	6	0
FcGI	10	5	10
FCGI	20	10	20
FCGS	24	12	30
FTPh	-28 -56	-18 -36	24 48

[8] Rappelons que la distance correspond au nombre de syllabes à parcourir dans la phrase pour relier le tronçon qui suit la pause à l'unité régie ou régissante.

[9] Fréquemment réduites y compris en situation de lecture.

3.4. Les paramètres acoustiques résultants

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

Le cheval trottait.

Elle comporte un seul groupe de souffle et 5 noyaux syllabiques. Sa déclinaison résultante est donc de –2,33 (*cf. supra*, § 3.1). Les lignes de déclinaison sont alors interpolées de façon à attribuer une valeur à chaque syllabe, avec comme pivot le milieu du groupe de souffle :

(5, 0, 0) (2, 0, 0) (0, 0, 0) (-2, 0, 0) (-5, 0, 0)

La 1ère syllabe est étiquetée IMG, la 3e FcGI, la 5e FTPh et les 2 autres IML. Les valeurs correspondantes de la table 5 sont extraites :

(-6, -6, -8) (0, 0, 0) (10, 5, 10) (0, 0, 0) (-28, -18, 24) (-56, -36, 48)

Elles sont ensuite superposées à la déclinaison :

(-1, -6, -8) (2, 0, 0) (10, 5, 10) (-2, 0, 0) (-33, -18, 24) (-61, -36, 48)

Enfin, la pause finale (10 unités) est insérée.

L'utilisation des paramètres acoustiques ainsi obtenus se fait au niveau du module acoustico-phonétique, pendant la production de la parole. La prosodie est appliquée au signal en cours de fabrication par l'interpolation et l'interprétation en temps réel des paramètres acoustiques en termes de variations par rapport aux propriétés intrinsèques des diphones[10].

4. CONCLUSION

Nous avons proposé ici un modèle qualitatif – phonologique – et quantitatif – acoustico-phonétique – tels qu'ils sont implémentés dans le module de génération automatique de la prosodie dans le système de synthèse à partir du texte Kali. L'approche repose sur l'exploitation de la congruence intonosyntaxique d'une part, sur la prise en compte d'indices typographiques et textuels d'autre part. Si Kali, par l'excellence de son intelligibilité donne entièrement satisfaction à nos utilisateurs aveugles, son acceptabilité reste néanmoins moyenne pour le grand public, la parole paraissant trop saccadée, voire artificielle. En pratique : la prosodie est jugée relativement plate en dehors des proéminences locales, malgré la déclinaison de la fréquence fondamentale. Aussi, des variations d'intensité et de vitesse de phonation vont-elles être introduites dans la déclinaison. Un complément d'analyse prosodique de corpus lus devrait confirmer la validité de la démarche et nous amener à une meilleure gestion du rythme dans son ensemble, une modélisation des patrons prosodique dans les parties non accentuées, une meilleure interpolation des paramètres acoustiques et une implémentation plus robuste des glissandos vocaliques.

[10] Pour plus de détails sur le module acoustico-phonétique, voir [Mor02].

BIBLIOGRAPHIE

[Abn92] Abney S. (1992), "Prosodic Structure, Performance Structure and Phrase Structure", *Proceedings, Speech and Natural Language Workshop,* San Mateo, Morgan Kaufmann Publishers, CA, pp. 425-428.

[Bou97] Boula de Mareüil P. (1997), *Etude linguistique appliquée à la synthèse de la parole à partir du texte,* Thèse de Doctorat, Paris XI.

[Hir93] Hirst D., Espesser R. (1993), "Automatic Modelling of Fundamental Frequency Using a Quadratic Spline Function", *Travaux de l'Institut de Phonétique d'Aix-en-provence,* 15, pp. 75-85.

[Lac99] Lacheret-Dujour A., Beaugendre B. (1999), *La prosodie du français,* Paris, Editions du CNRS.

[Lac01] Lacheret-Dujour A., Morel M. (2001), "Génération automatique de la prosodie pour la synthèse à partir du texte : le système Kali", actes des *Journées d'étude sur la prosodie*, Grenoble, V. Aubergé & A. Lacheret (éd.), à paraître.

[Mer01] Mertens P., Auchlin A., Goldman J.P., Grobet A. (2001), "L'intonation du discours : une implémentation par balises ; motifs et premiers résultats", actes des *Journées d'étude sur la prosodie*, Grenoble, V. Aubergé & A. Lacheret (éd.), à paraître.

[Mor02] Morel M., Lacheret-Dujour A. (2002), "Le logiciel de synthèse vocale Kali : de la conception à la mise en œuvre", *TAL,* Ch. D'Alessandro (éd.), Paris, Hermès, pp. 115-144.

[Pas90] Pasdeloup V. (1990), *Modèle de règles rythmiques du français appliqué à la synthèse de la parole,* Thèse de Doctorat, Aix-en-Provence.

[Ros99] Rossi M. (1999), *L'intonation, le système du français,* Paris, Ophrys.

[Slu93] Sluijter A., Terken J.M.B. (1993), "Beyond Sentence Prosody : Paragraph Intonation in Dutch", *Phonetica 50,* pp. 180-188.

[Van99] Vannier G. (1999), *Etude des contributions des structures textuelles et syntaxiques pour la prosodie : application à un système de synthèse vocale à partir du texte, Thèse de Doctorat, Université de Caen.*

[Ver99] Vergne J. (1999), *Etude et modélisation de la syntaxe des langues à l'aide de l'ordinateur :*

analyse syntaxique automatique non combinatoire, Diplôme d'habilitation à diriger des recherches, Université de Caen.