”Kali”, synthèse vocale à partir du texte - De la conception à la mise en oeuvre

Michel Morel, Anne Lacheret-Dujour

« Kali », synthèse vocale à partir du texte : de la conception à la mise en œuvre

Michel Morel, Anne Lacheret-Dujour

Laboratoire CRISCO, Université de Caen, 14032 Caen cedex
{morel, lacheret}@crisco.unicaen.fr

RESUME.

Résultat d'une collaboration universitaire et industrielle, Kali, logiciel de synthèse vocale à partir du texte en français, a été conçu pour les déficients visuels. Le texte fourni en entrée traverse successivement cinq modules (prétraitement, analyse syntaxique, génération de la prosodie, phonémisation, traitement acoustico-phonétique) avant d'être prononcé. Sa grande qualité est l'intelligibilité à vitesse d'élocution élevée.
Nous présentons dans cet article l'architecture générale du système et les principes retenus pour le développement des différents modules appelés en vue de la génération du signal synthétique.

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 five 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, the general architecture of the system and the main principles which were carried out for the development of the different units called for generating the speech signal are presented.

MOTS CLES :

Synthèse vocale, syntaxe, prosodie, transcription graphème-phonème, signal vocal, aveugle.

KEYWORDS :

Text-to-speech synthesis, syntax, prosody, text-to-phonem conversion, speech signal, blind.

Traitement Automatique des langues, volume 42 - N° 1 /2001, pages

1. Introduction

Nos recherches en synthèse vocale ont été initiées en 1981 avec la mise au point et la commercialisation d'un appareil portable à clavier phonémique destiné au handicap vocal [MOR 81]. En 1986, Synthé 3, système de synthèse par diphones doté d'un module de transcription graphème-phonème, lui succédait pour devenir en France, jusqu'au début des années 90, une référence à la fois dans le domaine du handicap vocal et dans celui du handicap visuel.

Dans la continuité de ces travaux, le projet de recherche appliquée Kali démarrait en 1995, réunissant deux laboratoires universitaires et deux partenaires privés[1]. Motivé par des besoins applicatifs précis mais également conduit dans le cadre de la recherche et de l'enseignement en phonétique-phonologie des langues, le programme Kali devait répondre à un triple objectif : disposer à terme d'une synthèse multilingue et, pour chaque langue, d'une voix non seulement intelligible, mais également acceptable pour l'utilisateur, enfin mettre en place une plate-forme de travail pour nos recherches en parole. L'action menée conjointement dans le domaine de la linguistique et de l'informatique a conduit au développement d'outils informatiques pour le traitement des données linguistiques (dictionnaires et règles pour l'analyse syntaxique et le calcul de la prosodie, règles de transcription graphème-phonème, outils d'écriture et de maintenance des règles) et la génération du signal de parole (analyse du signal vocal, mise au point de bases de diphones)[2]. Doté de deux voix masculines et d'une voix féminine, pour l'heure, Kali fonctionne en langue française[3]. Utilisable par la plupart des logiciels spécialisés pour déficients visuels[4], il peut également être intégré dans les matériels dédiés utilisés pour le handicap vocal.

Cet article fournit une présentation critique de l'architecture du système dans son ensemble, c'est-à-dire des différents modules appelés en vue de la génération du signal de parole, des traitements symboliques à l'analyse acoustico-phonétique. En nous appuyant sur divers exemples linguistiques, nous précisons les principes fondamentaux qui régissent l'organisation interne des modules présentés et la façon dont l'information circule d'un module à un autre.

---

1 Les laboratoires de linguistique et d'informatique de l'université de Caen (respectivement l'Elsap et le Greyc), la société d'informatique Electrel et l'association d'aveugles Club Micro son.

2 Après trois ans d'études menées grâce à un financement européen Feder, Kali a été commercialisé en juin 1999 et s'est fait remarquer par deux faits majeurs : l'obtention d'un prix ADER en recherche appliquée au printemps 1999 et la remise à M. Morel du CRISTAL CNRS en septembre 2000.

3 Une démonstration interactive est proposée sur le site : www.crisco.unicaen.fr

4 Ces logiciels permettent aux mal voyants d'avoir accès à l'outil informatique en décrivant oralement les opérations effectuées par l'ordinateur (accès vocal au contenu de l'écran ou à un document imprimé) ; voir notamment le système de lecture optique OpenBook Ruby (Arkenstone), le lecteur d'écran Jaws (Henter Joyce) et le logiciel d'agrandissement de caractères ZoomText.

2. Schéma de principe

Le système s'articule autour de cinq modules de traitement organisés séquentiellement (fig.1).

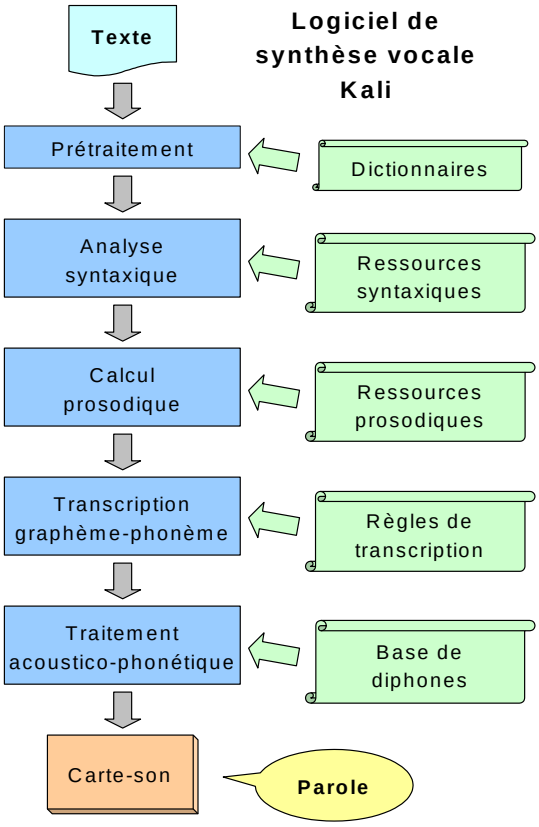


Figure 1. *Architecture de Kali*
(Occupation mémoire : 1 Mo + 6 Mo par voix)

Les quatre premiers modules reposent sur une exploitation déclarative des connaissances, la base de diphones, contenue dans le dernier module, est générée par un logiciel interactif de développement. L'objectif à court terme d'une synthèse bilingue (français-anglais) explique ce choix : une présentation déclarative des ressources linguistiques permet de passer rapidement d'une langue à une autre, seules les données, les dictionnaires et les règles changent, les outils de traitement restant les mêmes. Par ailleurs, une telle architecture facilite indéniablement la

lisibilité, le traçage, la mise à jour et la correction des règles développées au sein de chaque module, donc la maintenance générale du système. Nos règles déclaratives et contextuelles, qui opèrent sur des unités de différents niveaux (chaîne graphémique, mot, constituant syntaxique ou prosodique, phrase, paragraphe), se composent classiquement de deux parties, la première correspond aux conditions d'application de la règle et la seconde spécifie la structure à engendrer lorsque ces conditions sont réunies.
Afin d'illustrer ce fonctionnement modulaire et les critères utilisés pour constituer les ressources associées aux différents modules, nous nous proposons de suivre séquentiellement le cheminement d'une phrase exemple, depuis le prétraitement jusqu'au module acoustico-phonétique. Cette phrase, construite de façon à éclairer le traitement de plusieurs problèmes typiques de la synthèse à partir du texte, est la suivante :

(p) *L'ADN des hommes célèbres intéresse le président.*

3. Prétraitement du texte

En premier lieu, cette opération a pour objet de nettoyer le texte, c'est-à-dire de réduire le jeu de caractères à manipuler afin de faciliter le travail des modules subséquents. Elle est fondamentalement liée au problème de resyllabation et s'effectue par la consultation d'un dictionnaire qui compte 430 entrées. Un deuxième dictionnaire est ensuite activable de façon facultative par l'utilisateur afin de transcrire correctement les mots pour lesquels la transcription graphème-phonème par règles, trop généralisante, ne peut s'appliquer.
Le problème de resyllabation se pose essentiellement pour les sigles, les abréviations, les chiffres romains et les symboles, qui nécessitent une expansion préalable nécessaire au traitement prosodique pour que celui-ci puisse positionner correctement les accents : par exemple, le sigle *ADN* sans prétraitement serait considéré comme un mot d'une syllabe (/αδν/) et donnerait lieu à une accentuation erronée sur la première voyelle. Certes, ce problème ne se rencontre pas dans les architectures où la transcription graphème-phonème précède la génération de la prosodie. Cette stratégie souvent utilisée [HAM 89] [AUB 91] [BOU 97] présente un avantage indéniable pour la syllabation. Encore faut-il conserver un alignement mot à mot du texte alphabétique et du texte phonémique pendant le calcul de la prosodie, tout en disposant intégralement du marquage syntaxique. Bref, aucune solution n'est parfaite et, compte tenu de cet inconvénient, nous avons opté pour l'architecture actuelle au prix d'un prétraitement plus lourd.

Illustrons le principe général de la démarche par quelques exemples :
- Les sigles formés de consonnes (ex : *DST*) ne nécessitent pas de prétraitement particulier, puisque chaque consonne graphémique donne lieu à l'occurrence d'une syllabe phonémique (/δε/Ɛσ/τε/). Autrement dit, au niveau prosodique, le calcul accentuel ne demande pas de traitement intermédiaire et, lors de la transcription

graphème-phonème, il suffit d'activer les règles d'épellation correspondantes. En revanche, comme nous l'avons vu avec *ADN*, les sigles contenant des voyelles graphémiques reposent sur une resyllabation préalable au calcul accentuel. Dans notre exemple, la ligne suivante du dictionnaire s'applique :

ADN = a.d.n.

La notation *a.d.n.* étant reconnue comme standard des sigles par les autres modules de traitement (y compris les points, intégrés dans le sigle), le mot sera correctement épelé et accentué.
- Le traitement des abréviations, des symboles et des chiffres romains repose sur les mêmes principes. Ainsi :

Mlle = mademoiselle
$ = dollar
IVe = 4ème

Si la distinction majuscules/minuscules qu'illustre le troisième exemple, permet d'éviter l'homographie avec le mot *ive*, en revanche, certains symboles restent particulièrement ambigus *(film X, chapitre X)*. Dans de tels contextes, où la levée de l'ambiguïté est en définitive du ressort de la sémantique, le caractère est traité par le module de transcription graphème-phonème, en utilisant des heuristiques locales (/ικσ/ si précédé de *film vs.* /δισ/ si précédé de *chapitre*, *charles*, etc.).
A ce niveau de l'analyse, les majuscules, n'étant plus nécessaires, sont converties en minuscules afin de faciliter la suite du traitement.
Enfin, la transcription graphème-phonème n'étant jamais parfaite (présence dans les textes de mots spécialisés, de néologismes et d'emprunts à d'autres langues), un dictionnaire de réécriture peut être créé par l'utilisateur lui-même pour pallier les manques de la transcription par règles. Ce dictionnaire est fourni avec quelques entrées seulement, servant ainsi de modèle pour la création de nouvelles lignes :

zacharie = zakarie

A l'issue du prétraitement, notre phrase (p) devient :

(p1) *l'a.d.n. des hommes célèbres intéresse le président.*

4. Analyse syntaxique

L'analyse syntaxique constitue aujourd'hui une étape fondamentale en synthèse à partir du texte, aussi bien pour résoudre certains problèmes liés à la phonémisation (traitement des liaisons entre mots, résolution des ambiguïtés dues par exemple à la présence d'homographes hétérophones dans un texte) que pour construire le modèle prosodique. Du point de vue de la phonémisation, certes la mise en place d'heuristiques locales au sein même du module de transcription reste toujours une solution, mais ceci à condition de multiplier les règles qui, de toutes façons, ne pourront lever toutes les ambiguïtés. A l'inverse, la reconnaissance et le traitement

syntaxique correct des homographes hétérophones se résument la plupart du temps à effectuer une distinction simple entre deux catégories syntaxiques (nom/verbe ou adjectif/verbe par exemple). L'analyse syntaxique constitue donc un traitement plus économique et plus robuste. Concernant la génération automatique de la prosodie, si les structures intonative et syntaxique ne sont pas nécessairement congruentes, il n'en reste pas moins qu'elles sont associées, dans la mesure bien sûr où l'intonation respecte par ailleurs un jeu de contraintes rythmiques fondamentales [ROS 97] [LAC 99]. La segmentation en constituants syntaxiques (ici appelés « tronçons ») représente donc un point d'ancrage précieux pour poser une structure prosodique de base (cf. *infra* 4.1.). Nous reprenons pour notre analyse les principes formulés dans le cadre des grammaires de dépendance [GIG 97], 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. L'analyse s'effectue en trois phases : la segmentation en phrases et en mots, l'étiquetage de ceux-ci et leur regroupement sous forme de tronçons, 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.
Ces différents traitements sont réalisés par l'utilisation conjointe de dictionnaires et de règles appelés séquentiellement. Enfin, des marqueurs sont générés pour préparer le traitement des liaisons et celui des homographes hétérophones, effectués par le module de transcription.

4.1. Segmentation en phrases et en mots

Les unités à isoler sont respectivement les paragraphes, les phrases à l'intérieur des paragraphes et les mots dans les phrases. Certes le paragraphe n'est pas nécessaire à l'analyse syntaxique, mais il représente une unité de traitement prosodique à considérer (cf. *infra* 4.1.) : un saut de ligne indique donc le passage à un nouveau paragraphe. La phrase est repérée par un signe de ponctuation terminale (. ! ?), à une exception près pour le point : plusieurs points encadrant un groupe de lettres successifs sont associés à la présence d'un sigle.
Si la segmentation en paragraphes et en phrases ne présente pas de difficulté particulière, il en va autrement pour l'unité « mot ». D'une façon très générale, on peut affirmer que le mot se définit comme une chaîne de caractères compris entre deux blancs ou entre un blanc et un signe de ponctuation. Notre principe de segmentation est donc le suivant : les caractères de ponctuation, les tirets, parenthèses, guillemets et leurs variantes servent de séparateurs de mots et sont eux-mêmes considérés comme des mots[5]. L'apostrophe (dans le cas du français) est rattachée au mot qui la précède. A première vue donc, le repérage des mots qui composent une phrase semble simple. Néanmoins, on ne peut pas négliger l'ambiguïté réelle de certains caractères comme l'apostrophe et le tiret (la chaîne *aujourd'hui*

5 Au sens typographique du terme.

correspond à un mot, *j'arrive* est formé de deux mots, ibid. pour *porte-monnaie vs. voulez-vous*). En conséquence, les critères typographiques sont nécessaires mais non suffisants. Ce point explique en partie l'utilisation de dictionnaires.

4.2. Étiquetage des mots

Pour ce traitement, dénommé également « catégorisation », sont appelés quatre dictionnaires, des règles de déduction locale, enfin une base morphologique simplifiée.

Le **premier dictionnaire** regroupe des mots qui avaient été séparés par le premier traitement. Il recense et étiquette ainsi **190 formes graphémiques ambiguës** telles que les mots composés séparés le cas échéant par un tiret qui ne peut, dans ce contexte, être considéré comme un mot et comme un séparateur de mot. Pour les entrées polycatégorielles (ex : *a priori*, adverbe ou nom), les catégories possibles sont indiquées à la suite, l'ambiguïté étant levée plus tard au niveau de la déduction contextuelle.

Dans le **deuxième dictionnaire** sont recensés les **100 homographes hétérophones** traités actuellement par l'analyseur. Les formes dont la prononciation dépend de contraintes sémantiques (*fils, jet*, etc.) sont ici ignorées. Ces dernières seront traitées par le phonétiseur dans quelques contextes typiques (*fils de fer, jet privé*, etc.). Dans notre phrase exemple (p1), le mot *président* s'aligne sur une des entrées du dictionnaire :

président = (nom masculin singulier) ou (verbe pluriel 3ème personne)

Les **dictionnaires** associés à la consultation des **mots grammaticaux** et des **verbes** constituent les piliers de l'analyse. Les mots grammaticaux[6], peu nombreux et très stables, organisent pour une large part les relations entre les constituants syntaxiques, en outre, ils permettent de lever un grand nombre d'ambiguïtés. Soit la phrase :

(1) *Le musicien soufflait sans trêve dans ses bazouks et ses strapons. (B. Vian)*

Les déterminants permettent ici de catégoriser comme noms les néologismes *bazouks* et *strapons*. Précisons que la nature souvent polycatégorielle de ces entrées, nous amène, comme pour les mots composés, à proposer plusieurs étiquettes grammaticales, filtrées ultérieurement par l'étape de déduction contextuelle. Les verbes, quant à eux, servent de point d'ancrage à la segmentation en tronçons : toute séquence qui ne se construit pas autour d'une forme verbale est considérée par défaut comme une séquence nominale. D'où la nécessité d'avoir repéré correctement les verbes au préalable. En outre, les formes verbales, beaucoup moins nombreuses que les formes nominales et plus stables (moins de néologismes et d'emprunts) justifient

6 Cinq catégories de mots grammaticaux sont définies (les déterminants, les pronoms, les prépositions, les conjonctions et les adverbes s'ils ne dérivent pas d'un adjectif suivi du suffixe *-ment : fièrement, simplement*).

ce choix : les parties du discours sujettes à de fortes variations ne peuvent bien évidemment pas faire l'objet d'un codage lexical. Pour autant, il est impossible de lister toutes les formes verbales dans le dictionnaire, étant donné le nombre non négligeable de terminaisons qui nous amènerait à multiplier les entrées (quelques dizaines de milliers). Une solution plus judicieuse consiste à explorer simultanément une base de terminaisons (59 groupes différents, étant donné les nombreux verbes irréguliers) et une base de racines verbales (2500 racines). La recherche s'effectue sur chaque mot non encore catégorisé, jusqu'à ce qu'une correspondance puisse être établie entre une terminaison et un radical. Pour le mot *intéresse* de notre phrase exemple, la terminaison *se* est identifiée (groupes *coudre* et *cuire*), mais le radical *intéres-* n'existe pas pour ces groupes. En revanche, la terminaison *-e* est répertoriée dans le groupe 1, qui contient le radical *intéress-* :

e = (verbe groupe 1) (singulier 3ème ou 1ère personne)
intéress = (verbe groupe 1)

Le mot est alors étiqueté verbe, singulier, 3ème ou 1ère personne.
A ce stade de l'analyse, certains mots ne sont pas encore étiquetés et d'autres le sont dans plusieurs catégories. C'est là qu'interviennent les **règles de déduction contextuelle**. Au nombre de 42, celles-ci utilisent les informations apportées par l'étiquetage déjà réalisé et les propagent par déduction. Par exemple, si un pronom sujet est suivi d'un mot étiqueté déterminant ou pronom objet, ce dernier est ré-étiqueté pronom objet. Ainsi, dans la phrase *il l'entraîna dehors*, la catégorie du *l'* est précisée par la règle de déduction suivante :

(pronom sujet) + ((déterminant) ou (pronom objet))
→ (pronom sujet) + (pronom objet)

Les catégories nouvellement attribuées peuvent à leur tour être utilisées pour d'autres déductions. En ce qui concerne la chaîne (p1), aucune déduction sur l'homographie n'est réalisée ; les règles de déduction contextuelle ne suffisent donc pas à terminer l'étiquetage. Un complément est alors apporté par une base de **suffixes** (600 entrées) fondée sur des indices statistiques. Un suffixe, en effet, est souvent porteur d'une information syntaxique [BOU 97], y compris pour les néologismes (*foultitude*, *trompage*, etc.), par exemple, la finale *-age* permet d'identifier un nom masculin singulier[7].
A l'issue de ces traitements, les mots non étiquetés sont considérés par défaut comme nominaux (voir *supra* le dictionnaire de verbes). L'étape suivante, la mise en relation, permet d'actualiser certaines étiquettes, en levant notamment des ambiguïtés de type nom/verbe.

7 Les finales adverbiales en *-ment* sont répertoriées dans cette base.

4.3. Segmentation en tronçons et mise en relation

Une fois l'étiquetage terminé, quelques déductions simples effectuées directement par le programme conduisent à regrouper les mots fortement liés syntaxiquement en tronçons (appelés également « chunks » [ABN 92]) – ex : déterminant + nom (+ adjectif) ; auxiliaire + verbe, etc. –. A ce stade, notre phrase exemple est segmentée en 4 tronçons :

(p2) *(l'a.d.n.) (des hommes célèbres) (intéresse) (le président.)*

La mise en relation des tronçons entre eux, très importante pour hiérarchiser les frontières prosodiques (cf. *infra* 4.1.), est une opération beaucoup plus complexe, qui fait appel à un jeu de 41 règles. Notre système dispose de différentes mémoires dédiées chacune à une relation syntaxique déterminée (ex : relation sujet-verbe, nom-complément de nom, etc.). Chacune de ces mémoires est gérée comme une pile : de gauche à droite de la chaîne à traiter, chaque nouveau tronçon inséré dans une mémoire est empilé et devient donc le premier candidat pour une éventuelle mise en relation. Un groupe qui n'est plus en attente d'autres groupes (parce que déjà relié) est effacé de la mémoire[8]. Ainsi pour notre phrase exemple, le premier tronçon nominal (*l'a.d.n.*) est mémorisé comme sujet possible, donc en attente d'un tronçon verbal, le deuxième (*des hommes célèbres*), est relié au premier (relation nom-complément de nom) et retiré de la mémoire. Le tronçon verbal (*intéresse*) valide *l'a.d.n.* comme sujet, celui-ci est donc effacé de la mémoire. Pour le quatrième tronçon (*le président*), construit autour d'un homographe, la règle appliquée est la suivante :

(verbal) + (homographe nom/verbe) → (verbal) + (nominal) (relié -1) {N}

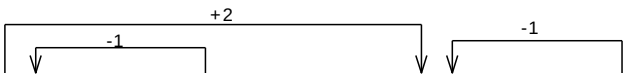
En d'autres termes, le tronçon homographe est redéfini comme nominal (adjonction du marqueur {N} à destination du phonétiseur) et relié au tronçon précédent. A l'issue de la mise en relation, la phrase est représentée de la façon suivante :

(p3) *(l'a.d.n.) (des hommes célèbres) (intéresse) (le {N}président.)*

8 Cette notion d'attente peut être rapprochée du concept de saturation valancielle chez Tesnière [TES 59].

4.4. Traitement des liaisons

La segmentation en tronçons délimitant des zones de forte cohérence syntaxique [LÉO 93], elle sert de point d'ancrage au traitement des liaisons. Une première règle simple peut, en effet, être posée : les liaisons sont obligatoires à l'intérieur des tronçons (*le**s** enfants*) et interdites d'un tronçon à l'autre *((conduit-le**s**) (en voiture))*. Quelques exceptions, cependant, méritent une attention particulière. D'une part, à l'intérieur d'un tronçon, certaines liaisons sont facultatives (nom + adjectif postposé : *des résultat**s** intéressants*), voire interdites (*un résulta**t** intéressant*). A l'inverse, la liaison peut être obligatoire d'un tronçon à l'autre *((il**s**) (arrivent), (elle**s**) (en veulent))*. Enfin, nombreux sont les contextes où la liaison est facultative malgré la segmentation en tronçons *((il) (mangeai**t**) (une pizza))*. Devant cette variation forte, quelle stratégie de marquage adopter ? La liaison obligatoire entre tronçons est résolue au niveau de l'analyse syntaxique par l'identification de deux contextes : il n'y a pas de coupure après un pronom sujet ou après un auxiliaire (*il**s** arrivent, ils son**t** amusants*). Les cas de liaison facultative seront résolus lors de la phonémisation (cf. *infra* 5.2.). En définitive, le rôle de l'analyseur syntaxique est ici de poser des marqueurs de coupure {C} interprétables par le module de transcription : dans un contexte de liaison possible, la présence du marqueur bloque son occurrence. La chaîne (p3) devient alors :



(p4) *(l'a.d.n.{C}) (de**s** hommes célèbre**s**{C}) (intéresse{C}) (le {N}président.)*

5. Calcul de la prosodie

La génération automatique de la prosodie suppose deux niveaux de représentation : (i) un niveau qualitatif, fondé sur une modélisation phonologique de l'intonation, consiste à générer un jeu de marqueurs abstraits pour rendre compte d'une structure prosodique hiérarchisée, (ii) le niveau quantitatif a pour fonction de faire correspondre à ces marqueurs les corrélats acoustico-phonétiques pertinents. L'ensemble du traitement (qualitatif et quantitatif) est effectué ici par un jeu de 90 règles, construites sur les bases de l'observation acoustique[9] de deux corpus de lecture oralisée [VAN 99], l'extrait d'un roman policier (520 mots) et un article de

9 Utilisation du logiciel Momel [HIR 99].

presse (500 mots)[10]. La structure prosodique 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 proéminence de syllabes accentuées, de différentes natures.
Ainsi, nous proposons une approche superpositionnelle de l'intonation[11], qui consiste à considérer les accents comme des proéminences locales subordonnées à l'intonation de groupes de souffle, de phrases et de paragraphes, elle-même modélisée par une ligne de déclinaison ou downdrift[12].

5.1. Analyse qualitative

Des constituants générés par l'analyse syntaxique, dérive une structure phonologique profonde formée d'une succession de groupes accentuels : hormis le pronom sujet toujours atone, tout constituant syntaxique donne lieu à la formation d'un groupe accentuel, défini comme une chaîne de syllabes dont la dernière est frappée par un accent démarcatif :

(2) *(Le présid<u>ent</u>) (parler<u>a</u>) (d<u>emain</u>)*

La question qui se pose est alors la suivante : comment dériver une structure intonative hiérarchisée et rythmiquement bien formée à partir de cette représentation accentuelle de base ? Nous développons l'hypothèse que la structure intonative s'articule autour de trois types de contraintes fondamentales : textuelles, syntaxiques et rythmiques. L'application de ces contraintes fait émerger six degrés de frontières intonatives.
Dans le détail, les **contraintes textuelles** nous amènent à manipuler trois unités de traitement : le paragraphe, la phrase et le groupe de souffle, ce dernier étant basé principalement sur la ponctuation interne à la phrase[13]. Ces trois unités se caractérisent par une déclinaison et une pause terminale de durée variable (la pause la plus forte est attribuée à l'unité 'paragraphe').

---

10 Le choix du corpus – texte et non phrases isolées – a été ici décisif pour vérifier l'hypothèse selon laquelle la dimension textuelle du message à synthétiser représente un paramètre essentiel dans la construction prosodique.

11 Voir Rossi [ROS 99] pour une présentation des différentes théories : superpositionnelle, linéaire, morphologique, hiérarchique.

12 Tendance de la fréquence fondamentale à décroître progressivement du début à la fin d'un énoncé.

13 Unité démarquée à sa droite par une virgule, un point-virgule ou deux-points.

D'où un premier jeu de frontières, hiérarchisées de la façon suivante :

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

Les **contraintes d'alignement syntaxique** dérivent du calcul des dépendances syntaxiques (contiguës ou à distance). En reprenant les notations utilisées dans la section précédente, nous illustrons la dépendance contiguë par l'exemple suivant :

(3) *(les enfants) (mangent) (leur soupe)*

Dans ce contexte, la relation de contiguïté syntaxique entre deux tronçons linéairement adjacents s'exprime sur le plan intonatif par une proéminence accentuelle associée à un allongement de la dernière syllabe pleine du premier tronçon.
Dans les contextes de dépendance à distance, l'élément régi et l'unité régissante n'entrent pas dans une relation linéaire de contiguïté, comme dans l'exemple suivant :

(4) *(les étudiants) (avaient appris) (en arrivant) (la triste nouvelle)*

où le circonstanciel *en arrivant* isole le complément d'objet de son régissant. Dans ce contexte, la relation de non-contiguïté syntaxique entre deux tronçons linéairement adjacents est marquée intonativement par une proéminence accentuelle associée à un allongement plus prononcé de la dernière syllabe pleine du premier tronçon et par l'insertion d'une pause dont la durée est proportionnelle au nombre de syllabes à parcourir dans la phrase pour relier l'unité régie à son régissant (ici une distance de 4 syllabes). Autrement dit, les deux degrés accentuels générés sont associés à un **principe de dominance intonative** qui traduit explicitement les deux types de dépendances syntaxiques (contiguës ou à distance). Deux nouvelles frontières sont ainsi définies :

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

En dernier ressort, la prise en compte des contraintes rythmiques consiste à insérer ou effacer certains marqueurs accentuels et pausals. En premier lieu, l'application d'un **principe de régulation temporelle** 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, par exemple :

(5) *(il promène) (ses enfants) (dans le jardin)*

(6) *(il promène) (les enfants) (de Nathalie) (et Vincent) # (dans le jardin)*

Dans l'exemple (5), l'allongement de la dernière syllabe de *ses enfants* reste important (FCGI), mais la pause (symbole '#') est effacée. Nous émettons alors

l'hypothèse que, lorsque la dépendance à distance est suffisamment longue pour justifier le maintien de la pause (6), il n'y a pas de différence fondamentale entre ce type de frontière et celui qui serait généré par l'existence d'une virgule typographique. Nous regroupons ainsi dans la même catégorie (FCGS) tous les groupes de souffle de taille inférieure à la phrase. Soit les règles suivantes pour illustration :

((tronçon) et (virgule)) → FCGS
(tronçon) + ((tronçon) et (distance[14] >= 8)) → FCGS
(tronçon) + ((tronçon) et (0 < distance < 8)) → FCGI
(tronçon) + ((tronçon) et (distance[15] = 0)) → FcGI

Considérons maintenant la **contrainte de régulation accentuelle** [PAS 90] selon laquelle un groupe constitué d'un nombre de syllabes inaccentuées trop important (4 syllabes ou plus) est accentué sur la première syllabe à attaque consonantique de son premier mot lexical *(une activité valorisante)*, et donne ainsi lieu à la formation d'un pied métrique[16]. Une dernière frontière est donc enfin définie :

*Niveau **6 : FPM*** *Frontière de Pied Métrique*

Deux mots enfin sur le **principe de non-collision accentuelle** [DEL 84] : dans la plupart des modèles, ce principe conduit à bloquer l'occurrence de deux accents terminaux contigus en effaçant le premier s'il marque un groupe monosyllabique :

(7) *(Paul) (dort)* → *(Paul dort)*

ou en le déplaçant à l'intérieur d'un groupe polysyllabique :

(8) *(Valérie) (dort)* → *(Valérie) (dort)* ou *(Valérie) (dort)*

Pour l'heure, nous n'avons pas jugé pertinent d'appliquer cette règle, posant comme hypothèse perceptive que la contiguïté accentuelle est acceptable dans la mesure ou les deux accents adjacents n'ont pas le même degré de proéminence (voir *supra :* le principe de dominance intonative). Une synthèse de la hiérarchie intonative ainsi définie est proposée tableau 1.

---

14 Où la distance correspond au nombre de syllabes qui séparent deux tronçons contractant une relation de dépendance syntaxique.

15 La distance est nulle quand les tronçons en relation sont linéairement adjacents (paramètres de relation +1 ou –1, cf. *supra*, phrase p4).

16 Unité prosodique caractérisée par un accent rythmique non terminal.

Niveau hié-rarchique	Déclinaison (F0)	Pause	Allongement	Proéminence (F0, intensité)
1	**FTPg**	**FTPg**	**FTPg**	
2	**FTPh**	**FTPh**	**FTPh**	
3	**FCGS**	**FCGS**	**FCGS**	**FCGS**
4			**FCGI**	**FCGI**
5			**FcGI**	**FcGI**
6				**FPM**

Tableau 1. *Hiérarchie des frontières intonatives et paramètres phonétiques associés (F0 = fréquence fondamentale)*

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

5.2. Analyse quantitative

L'analyse quantitative implique un choix préalable d'unités de mesures aussi pertinentes que possible. Nous avons choisi pour la fréquence fondamentale, l'intensité et l'allongement, des unités logarithmiques de granularité suffisamment fine pour assurer une bonne précision (tab. 2).

Afin de donner une idée de cette granularité, nous fournissons un équivalent tonal, étant entendu que pour l'intensité et l'allongement, il s'agit du rapport qui correspondrait à l'unité tonale pour la hauteur. Les pentes sont exprimées relativement au nombre de syllabes du groupe concerné, et les pauses en nombre de phonèmes (un phonème dure environ 100 ms à vitesse de phonation modérée).

Les variations logarithmiques des trois types de paramètres acoustiques sont calculées par rapport à des valeurs syllabiques de référence (les syllabes inaccentuées) implicitement contenues dans la base de diphones préenregistrée. Un jeu de paramètres phonétiques est défini et quantifié pour chacun des niveaux de la hiérarchie intonative, ainsi que pour les différentes modalités de phrase (déclarative, interrogative, exclamative, suspensive) et de groupe de souffle (pauses syntaxiques et/ou typographiques). Par exemple, la modalité interrogative est représentée par une proéminence terminale (F0 et allongement) et une déclinaison plus forte en valeur absolue que la modalité déclarative.

Paramètre acoustique	Rapport correspondant à un incrément de 1	Équivalent tonal
Fréquence fondamentale	1,00726	1/16^{e} ton
Intensité	1,0146	1/8^{e} ton
Allongement	1,0146	1/8^{e} ton

Tableau 2. *Unités utilisées pour les paramètres acoustiques*
Ces unités étant logarithmiques, un incrément de 1 correspond à un rapport (colonne 2) fonction de la granularité recherchée (ex : si la valeur de base 0 correspond à une fréquence de 100 Hz, la valeur 1 correspond à 100,726 Hz).

L'interprétation phonétique des objets posés par l'analyse qualitative consiste en premier lieu à calculer une ligne de déclinaison pour chacun des trois niveaux concernés (tab. 1). La mélodie de l'énoncé est ensuite synthétisée en superposant la partition intonative de chaque niveau. La réinitialisation subséquente à la déclinaison (remise à niveau de la fréquence fondamentale) n'est pas calculée mais dérive du calcul des pentes.
Précisons ici que la pente, sensiblement constante pour des groupes de souffle de taille réduite (< 5 syllabes), diminue en valeur absolue quand la taille des groupes augmente, de telle façon que l'amplitude maximale de variation ne dépasse jamais le registre d'un locuteur humain en situation de lecture. Autrement dit, plus le nombre de syllabes augmente, moins la déclinaison est marquée [BEA 94]. Le paramètre de pente *p*, dépendant du nombre de syllabes *s*, se décompose donc 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 :

$$p = \frac{A}{s + \frac{A}{P}} \qquad [1]$$

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

Le tableau 3 présente nos modèles de lignes de déclinaison pour la modalité déclarative et la virgule.

Niveau hiérarchique	Amplitude maximale	Pente maximale
1. FTPg	-2	-0,5
2. FTPh	-4	-1
3. FCGS	-10	-2

Tableau 3. *Paramètres des lignes de déclinaison (modalité déclarative)*
Dans notre approche superpositionnelle, la phrase isolée (p4) formée d'un seul groupe de souffle prend donc pour paramètres respectifs la somme des paramètres des 3 niveaux : -16 et -3,5.

La règle suivante illustre l'attribution des valeurs de déclinaison :

(niveau = FTPh) et (mode = déclaratif) → déclinaison (-4, -1)

Pour la modalité déclarative et la pause typographique et/ou syntaxique, nos modèles de pauses se déclinent comme suit (tab. 4) :

Niveau hiérarchique	Paramètre durée	Correspondance en ms
1. FTPg	4	400
2. FTPh	6	600
3. FCGS	distance[17] x 0,25 (maximum 6)	200 à 600

Tableau 4. *Paramètres des pauses (modalité déclarative)*
Les pauses des niveaux 1 et 2 s'ajoutent, ce qui veut dire que la dernière phrase d'un paragraphe est suivie d'une pause de 10 unités (soit 1 s à vitesse de phonation modérée) contre 6 unités pour une phrase située à l'intérieur d'un paragraphe. Par définition, la pause de niveau 3 ne peut pas coexister avec les deux précédentes (frontière continuative et non terminale).

Le calcul acoustico-phonétique des proéminences locales distingue quatre paradigmes syllabiques :

- les syllabes inaccentuées des mots lexicaux (IML),
- les syllabes inaccentuées des mots grammaticaux (IMG)[18],
- les syllabes accentuées démarcatives de pieds métriques (FPM),
- les syllabes accentuées finales de groupes.

---

17 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.

18 Dont les syllabes sont fréquemment réduites y compris en lecture oralisée.

Nos principaux paramètres acoustiques sont donc les suivants pour la modalité déclarative (tab. 5) :

Type de syllabe	F0	Intensité	Allongement
IML	0	0	0
IMG	-6	-6	-8
FPM	12	8	0
FcGI	18	10	18
FCGI	24	10	24
FCGS	32	0	32
FTPh	-28 -56	-18 -36	20 40

Tableau 5. *Paramètres acoustiques en modalité déclarative*
La dernière syllabe de la phrase possède deux valeurs afin d'opérer un glissando entre le début et la fin de la voyelle.

La règle suivante illustre l'attribution des paramètres acoustiques :

(syllabe = FcGI) et (mode = déclaratif) → param acou (18, 10, 18)

A la fin de l'analyse quantitative, il reste à concrétiser le modèle en plaçant, sur chaque noyau syllabique préalablement identifié, les marqueurs nécessaires au module acoustico-phonétique qui se chargera de les interpoler plus finement, phonème par phonème. Les lignes de déclinaison sont donc superposées et interpolées de façon à attribuer une valeur à chaque syllabe, les contextes de proéminence et d'allongement spécifiés et les pauses finales de groupes insérées. Le début de la phrase (p4) peut alors être représenté de la façon suivante :

(p5) *(6, 0, 0)l'a.(5, 0, 0)d.(22, 10, 18)n.{C} (-3, -6, -8)des (2, 0, 0)hommes (…)*

Avec 14 syllabes, une pente limite de -3,5 et une amplitude limite de -16, l'amplitude de la déclinaison est de -12 selon la formule [1], d'où des valeurs allant de 6 à -6 (6 à 2 pour le début de la phrase, paramètre de gauche de chaque triplet). Sur la troisième syllabe, une proéminence de 18 unités superposée à la déclinaison porte la modification de F0 à +22. A l'inverse, la quatrième syllabe (mot grammatical) subit une réduction vocalique elle aussi superposée à la déclinaison.

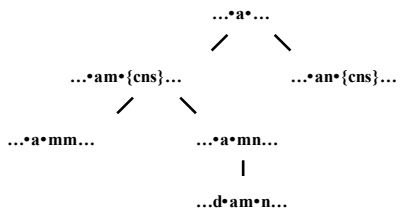
6. Transcription graphème-phonème

Notre module de transcription graphème-phonème est fondé sur un jeu de règles hiérarchisées, organisées sous forme d'une liste arborescente [MOR 98]. Considérons par exemple l'extrait des règles suivant :

…•a•…	**α**	
– |…•am•{cns}…	**A)⌉**	(ample, chambre)
– – |…•a•mm…	**α**	(gamme)
– – |…•a•mn…	**α**	(amnistie)
– – – |…d•am•n…	**α⌉**	(damné, condamner)
– |…•an•{cns}…	**A)⌉**	(manche, banque, grange)

où : •xxx• = chaîne à transcrire
{cns} = contexte graphémique consonantique
… = contexte quelconque

Chaque règle est écrite dans le contexte graphique où elle s'applique, en retrait d'un nombre d'indentations lié à son niveau hiérarchique. Des exemples sont fournis entre parenthèses pour illustrer une règle et la contrôler. La liste de règles ci-dessus peut être représentée par un arbre de dépendance, comme suit :



Où, chaque règle correspond à un nœud de l'arbre, ses exceptions étant associées aux nœuds qui lui sont connectés. Une telle structure optimise le parcours de l'interpréteur en éliminant les tests inutiles. Comparé aux systèmes actuels, dans lesquels les règles sont structurées sur une dimension, de la plus particulière à la plus générale [BEL 92] [BOU 97], la structure arborescente présente l'avantage de donner une place précise à chaque règle. Elle permet ainsi d'éviter la perte progressive de lisibilité au fur et à mesure de l'implémentation de nouvelles règles, tout en conservant une cohérence théorique aux sous-ensembles créés : en parcourant l'arbre, nous retrouvons les règles de base du français près de la racine (*ai*, *ain*, *am*, *an*, etc.), puis dans les niveaux suivants des règles plus particulières (consonnes finales muettes ou non, terminaisons en *-ent*, *h* disjonctif ou non, *i*-voyelle, *s* entre voyelles, *ti*-voyelle prononcé /**σι**/ ou /**τι**/, etc.), et aux derniers niveaux les règles les plus particulières.
La prononciation du français comporte plus de 1 000 règles élémentaires du type de celles citées. En ajoutant les mots d'emprunt et les noms propres les plus courants, on

arrive à plusieurs milliers de règles. Kali comporte actuellement 5 000 règles de transcription.

6.1. Traitement du *e* caduc

Notre traitement du 'e' caduc s'appuie sur la règle des trois consonnes [GRA 14], selon laquelle le 'e' précédé de deux consonnes phonémiques se prononce si le contexte droit est également consonantique (*garn**e**ment*). Il est effacé dans les autres contextes (*pomm**e**, fièr**e**ment*). Etant donné qu'il s'agit ici de tester des contextes phonémiques auxquels les règles n'ont pas accès (seuls l'environnement graphémique peut être défini au sein de la structure de règles), la règle principale génère un marqueur intermédiaire à destination de l'interpréteur :

…•e•… **{↔}** (pomme, fièrement, garnement)

Ainsi, pour la transcription du 'e' de *garn**e**ment*, l'interpréteur, ayant déjà transcrit la chaîne /**γαρν**/, place le marqueur dans un tampon en attente de la transcription du phonème qu'il précède. Celui-ci correspondant bien à un phonème consonantique (/**μ**/), l'interpréteur teste le tampon, le vide et complète la chaîne phonémique /**γαρν↔μ**/.
Une règle complémentaire a, par ailleurs, été définie pour transcrire correctement les 'e' des syllabes initiales de mots (*p**e**lage vs. plage, r**e**faire, j**e***). Enfin, d'autres règles, plus particulières, sont écrites dans le fichier de règles, comme dans l'exemple suivant où le schwa est maintenu malgré la présence d'une seule consonne devant lui :

– – – |soup•e•s{voy}… **↔** (soupeser)

6.2. Traitement de la liaison

Le module de transcription graphème-phonème définit les liaisons potentiellement obligatoires, principalement derrière les auxiliaires, les verbes avec inversion du sujet, les déterminants, les pronoms personnels sujets et les adjectifs antéposables :

– – – – |célèbre•s• **[ζ]** (célèbres idées)

Le phonème de liaison est placé entre crochets. Lorsque l'interpréteur le rencontre, il le place dans un tampon qui sera validé si le phonème suivant est une voyelle.
La rencontre d'un marqueur de coupure {C} ou d'une pause, en provenance de l'analyseur syntaxique, efface la liaison proposée. Ainsi, dans la phrase (p5), le premier marqueur {C} rencontré est sans action car aucune liaison n'est proposée. En revanche, le deuxième marqueur efface le tampon de liaison **[ζ]** créé à la lecture du *s* de l'adjectif *célèbres*.
Des liaisons obligatoires sont également implémentées dans les mots composés (*Etat**s**-unis, de par**t** et d'autre*). Pour le traitement des liaisons facultatives, nous

avons opté pour une stratégie minimaliste qui consiste à ne pas générer la liaison entre un nom et un adjectif postposé ou un verbe et son complément sauf dans les contextes de figement lexical (*des personne**s** âgées*).

6.3. Traitement de l'homographie

La majorité des homographes hétérophones (une centaine) repose sur l'ambiguïté verbe/nom ou adjectif (terminaisons *-ent, -tions, -er*). L'analyseur syntaxique a déjà défini la catégorie de ces derniers (taux de réussite supérieur à 95 %) et a placé un marqueur spécifique {V} ou {N} (verbal ou nominal). Il reste maintenant à tester l'un ou l'autre de ces marqueurs dans une exception à la règle de base :

– – – – – |**présid•ent•**	**A)**	(le président)
– – – – – – |{V}**présid•ent•**		(ils président)

Pour les homographes hétérophones ne reposant pas sur une ambiguïté de catégorie, la désambiguïsation dépend de critères sémantiques difficiles à modéliser. Nous répertorions alors des contextes typiques :

– – |**fi•ls•**	⌉σ	(mon fils)
– – – |**fi•ls• (à plomb+de fer)**	λ	(fils à plomb, fils de fer)

Le signe '+' correspond à l'opérateur 'ou' qui permet de juxtaposer une variété de contextes entre parenthèses. Plusieurs dizaines de contextes sont ainsi définis pour le mot *fils*.

6.4. Transmission des marqueurs prosodiques

Lorsque l'interpréteur rencontre des marqueurs prosodiques (F0, intensité, durée intrinsèque et extrinsèque), il les transmet à la chaîne phonémique en construction, permettant ainsi de conserver l'alignement de ceux-ci entre la chaîne alphabétique et la chaîne phonémique. Cette dernière possède alors toute l'information nécessaire à la production de la parole. Par exemple, le début de la phrase (p5) devient :

(p6) *(6, 0, 0)λα⌉(5, 0, 0)δε(22, 10, 18)Ev(-3, -6, -8)δε⌉ζ(2, 0, 0)□μ (…)*

7. Module acoustico-phonétique

Le principe de la synthèse par concaténation de diphones[19], qui produit une voix de qualité pour un coût de développement raisonnable [CHA 89] [HAM 89] [ALE 96] [DUT 97], a été ici retenu. Le moteur du module est le générateur de parole et sa res-

19 Portion de signal allant de la partie stable d'un phonème à la partie stable du phonème suivant.

source principale une base de diphones. Nous utilisons pour le français un jeu de 33 phonèmes et un silence, soit 34 x 34 = 1 156 diphones représentant toutes les combinaisons de deux phonèmes. La base de diphones est préalablement fabriquée, à l'aide d'un logiciel de traitement de la parole développé dans notre laboratoire, par prélèvement des diphones à partir de voix naturelles enregistrées. A chaque locuteur différent correspond une base de diphones spécifique, donnant au système de synthèse vocale la « voix » de ce locuteur. Un corpus oral destiné au prélèvement des diphones a été mis en place pour chaque locuteur, l'enregistrement étant effectué à la fréquence d'échantillonnage de 22050 Hz, valeur la plus utilisée actuellement pour les applications multimédia de bonne qualité.

L'agencement des diphones étant différent pour chaque message à synthétiser, il est nécessaire que ceux-ci soient normalisés pour éviter les discontinuités de fréquence fondamentale, d'intensité, de durée, de timbre et de phase. Il faut donc établir une norme portant sur chacun de ces paramètres, pour chaque phonème interface entre diphones. Sur un jeu de 34 x 34 diphones, un même phonème est représenté 68 fois à l'une ou l'autre extrémité des diphones. En ce qui concerne la fréquence fondamentale et l'intensité, la moyenne de ces 68 occurrences donne une valeur assez précise que nous pouvons considérer comme la norme du phonème considéré pour le locuteur choisi. La normalisation consiste alors à modifier les extrémités des diphones pour les faire correspondre à cette norme, tout en interpolant les valeurs intermédiaires. La durée des diphones ne peut pas être calculée de cette façon, car chaque occurrence de diphone est un exemplaire unique, d'où une grande dispersion des valeurs de durée. Nous avons donc construit un modèle de durée dépendant de la catégorie de chacun des deux phonèmes constituant le diphone (voyelle, glissante, fricative sourde, fricative voisée, etc.) à partir de mesures des durées dans des parties non accentuées de corpus oraux. Le diphone est alors allongé ou raccourci pour coïncider avec la durée modèle. La normalisation de timbre et de phase, plus complexe et imparfaite, peut se faire par analyse spectrale-modification-resynthèse ou, comme dans notre cas, par un mélange progressif dans le domaine temporel aux extrémités du diphone.

Les traitements que nous venons d'évoquer sont pour la plupart fondés sur le phénomène de quasi-périodicité du signal vocal dans ses parties voisées. Ils nécessitent une synchronisation des objets de calcul sur les périodes de ces parties et impliquent de ce fait un repérage préalable de celles-ci.

Nous détaillons dans cette section les principes retenus pour le repérage des périodes, la modification de la fréquence fondamentale et de la durée, la normalisation du timbre et de la phase, et montrons, à partir de notre exemple de phrase, comment le module acoustico-phonétique produit le signal vocal synthétique.

7.1. Repérage des périodes

Dans le domaine de la synthèse vocale, chaque laboratoire possède son algorithme de repérage des périodes, qui se compose généralement de deux opérations

principales : filtrage passe-bas et auto-corrélation. Le filtrage (à 400 Hz pour notre part) est destiné à éliminer le bruit des parties sourdes et les artéfacts non périodiques, pour ne conserver que les premiers harmoniques des parties voisées. Ensuite, le signal est auto-corrélé avec un décalage variable, jusqu'à un maximum correspondant théoriquement à la période T. Or cette méthode présente des maxima non seulement pour un décalage de T, mais aussi de 2T, 3T, etc. voire de T/2, d'où un taux d'erreur non négligeable. La réduction de ces erreurs demande une expertise beaucoup plus complexe, encore loin des capacités humaines.
Comme nous voulions un repérage sans erreur, nous avons adopté une méthode semi-automatique en trois étapes : (i) repérage des maxima relatifs d'intensité (un ou plusieurs par phonème voisé) et marquage d'une période pour chaque maximum, (ii) propagation des marques de périodes par itération de part et d'autre du maximum, (iii) marquage des parties non voisées par interpolation. L'intervention humaine est ainsi optimisée : après l'étape (i), subsiste environ une erreur toutes les 10 s, dont la correction consiste à modifier seulement une ou deux marques (fig. 2).

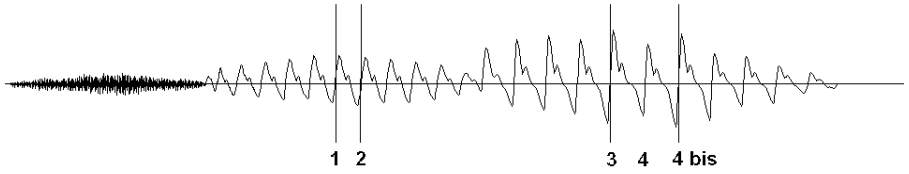


Figure 2. *Première étape du repérage des périodes*
Chaque marque est placée sur la plus grande pente détectée localement. L'intervention humaine va permettre de déplacer la marque (4 bis), erronée, jusqu'à sa position correcte (4).

Ensuite, l'étape (ii) bénéficie des marques déjà connues et permet ainsi d'ajouter au score de corrélation celui de la comparaison des durées de deux périodes successives. En clair, il est impossible que la périodicité passe brutalement au double ou à la moitié de la valeur précédente. Les rares erreurs se situent aux frontières des parties voisées, c'est-à-dire aux endroits où l'algorithme itératif doit être arrêté. L'étape (iii) a seulement pour fonction de segmenter les parties non voisées de façon à permettre des traitements du même type que ceux des parties voisées.

7.2. Modification de la fréquence fondamentale et de la durée

Les méthodes utilisées pour modifier la fréquence fondamentale et la durée en restant dans le domaine temporel dérivent généralement de la méthode TD-PSOLA, utilisée par le système Proverbe, du CNET [CHA 89]. Elles présentent le gros avantage de nécessiter peu de calculs, ce qui permet de les appliquer en temps réel pendant la production de la parole pour concrétiser la prise en compte des variations prosodiques, tout en conservant au mieux la qualité du timbre original.

En pratique, le signal de période courante T est découpé en signaux élémentaires de largeur 2T, résultant du produit du signal d'origine par une fenêtre en cloche, de largeur 2T, centrée sur la partie de plus forte énergie de chaque période. Si les fenêtres sont sinusoïdales (fenêtres de Hanning), la somme des signaux élémentaires, qui se recouvrent mutuellement sur la largeur d'une période, reproduit le signal d'origine (fig. 3). La forme des fenêtres diffère selon les auteurs. Après des essais perceptifs sur plusieurs types de fenêtres, nous avons constaté que c'était moins la forme de la fenêtre que sa position par rapport au signal qui était importante, et nous avons adopté la forme sinusoïdale, qui présente l'avantage d'une bonne répartition de l'énergie, permettant d'étendre le traitement des parties voisées du signal à celui des parties sourdes.

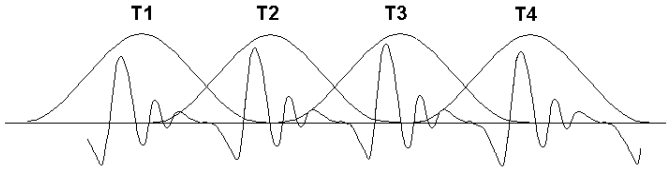


Figure 3. *Juxtaposition des signaux élémentaires après fenêtrage*

La diminution de la période consiste à resserrer les signaux élémentaires tout en diminuant conjointement la largeur des fenêtres sinusoïdales, de façon à conserver une répartition uniforme de l'énergie. L'augmentation de la période correspond à l'opération inverse, à ceci près qu'il n'est pas souhaitable de dilater les fenêtres, car, trop larges, elles contiendraient alors des traces non négligeables de la périodicité d'origine, créant des phénomènes de réverbération préjudiciables à la qualité de la parole. Les fenêtres sont donc écartées sans modification, ce qui ne détériore pas significativement la qualité de la voix synthétique. En revanche, les parties non ou peu voisées (en pratique inférieures à un certain taux de voisement) dont l'énergie est répartie plus uniformément, peuvent bénéficier de la dilatation des fenêtres, évitant ainsi un phénomène de périodisation, qui serait lui aussi préjudiciable à la qualité de la voix [ALL 77] [GRI 84].

La modification de durée (ou de son inverse, la vitesse de phonation) est effectuée en dupliquant ou en retirant des signaux élémentaires. Le problème de largeur des fenêtres ne se pose pas ici, les fenêtres à superposer étant de tailles égales ou voisines. En conséquence, la modification de durée par les méthodes temporelles affecte très peu la qualité du signal résultant.

L'évaluation perceptive de nos méthodes de modification nous a permis d'obtenir les résultats suivants (V = vitesse de phonation) :

- F0 ± 3 tons détérioration négligeable
- F0 ± 1 octave détérioration nette (acceptabilité moyenne)
- V x 2 ou V / 2 détérioration négligeable
- V x 3 ou V / 3 détérioration nette (acceptabilité moyenne)

Les variations maximales dues à la prosodie dépassent rarement 3 tons sur la fréquence fondamentale et un rapport 2 sur les durées. En définitive, nous pouvons souligner la performance de nos méthodes de modification, qui n'ont pas d'influence notable sur la qualité finale de la voix synthétique.

7.3. Normalisation du timbre et de la phase

Selon le contexte et compte tenu de la variabilité inhérente à un locuteur donné, le timbre d'un même phonème peut différer d'une occurrence à l'autre, créant une discontinuité spectrale lors de la concaténation des diphones (fig. 4). Il est donc nécessaire de lisser cette discontinuité.

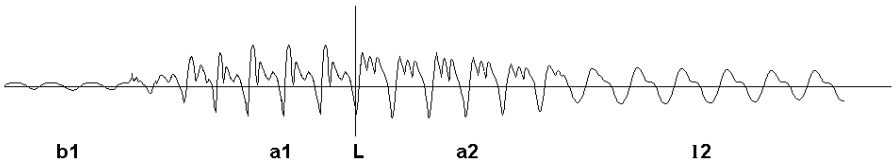


Figure 4. *Concaténation des diphones ba et al*
La discontinuité de timbre apparaît ici nettement au niveau de la limite L.

Le système MBROLA [DUT 97] utilise le principe d'analyse-modification-resynthèse des diphones. Schématiquement, l'analyse spectrale met en évidence un ensemble de formants propre à chaque phonème. Le principe consiste à infléchir ces formants au voisinage de chaque extrémité du diphone de façon à les faire coïncider avec la norme établie pour le phonème considéré, puis à resynthétiser le diphone à partir du nouveau spectre. Se pose alors le problème des phases de chaque harmonique, qui doivent coïncider lorsque les diphones sont juxtaposés, sous peine de discontinuités. Il faut donc normaliser également les phases des extrémités et les réajuster progressivement à l'intérieur du diphone, ce qui modifie sensiblement le spectre. La voix est plus lisse qu'avec les méthodes temporelles, mais d'un aspect un peu plus synthétique, bien que d'excellente qualité.
De son côté, le système Proverbe traite le problème dans le domaine temporel en établissant une zone de transition entre diphones adjacents et en effectuant un mélange progressif de ceux-ci en temps réel, par la méthode TD-PSOLA, au moment de la production de la parole. Les harmoniques et leurs phases glissent ainsi d'un phonème à l'autre sans discontinuité. Nous avons opté pour une méthode similaire, mais en l'effectuant au moment de la constitution de la base de diphones, ceci afin de diminuer le temps de calcul lors de la production de la parole. Nous procédons en deux étapes : (i) au moment du prélèvement, un tri des diphones candidats est effectué en fonction de critères d'acceptabilité auditive (nous comparons la qualité de mots fabriqués avec les divers candidats) et d'un score de distance spectrale permettant de minimiser les discontinuités de timbre ; (ii) un mélange progressif

(utilisant nos fenêtres sinusoïdales) est effectué aux extrémités du diphone avec un phonème type, déterminé lui aussi à l'aide d'un score de distance spectrale, et prélevé à l'extrémité du diphone qui le contient. A l'issue de ces traitements, les diphones peuvent être concaténés directement.

7.4. Production du signal vocal

Les diphones à concaténer sont choisis en lisant la chaîne phonémique, par exemple, pour la phrase (p6) :

#λ λα αδ δε εE Eν νδ δε εζ ζ□ □μ μσ σε ελ λE Eβ βρ (...)

où '#' = silence

Chaque diphone pointe sur un signal déjà normalisé, prêt à l'emploi. Restent à exploiter les paramètres morphologiques et acoustiques présents dans la chaîne phonémique pour doter la parole d'un certain relief. Les paramètres morphologiques (allongements et compressions pour le français) proviennent du module de transcription graphème-phonème. Dans le cas de la phrase 6, nous avons transcrit un allongement pour l'épellation du *a* (distinct de la transcription de *à*) et pour la voyelle de *des* (distinct de la transcription de *dé*). A ces paramètres morphologiques correspondent des paramètres acoustiques (ex : allongement de 10 unités pour '┐', soit environ 15 % de la durée de la voyelle) qui sont ajoutés aux paramètres acoustiques provenant du module de génération de la prosodie. De plus, les paramètres acoustiques sont étendus aux phonèmes qui n'en comportaient pas. Le début de la phrase (p6) peut alors être représenté de la façon suivante :

(p7) *(6, 0, 0)#(6, 0, 0)λ (6, 0, 0)λ(6, 0, 10)α (6, 0, 10)α(5, 0, 0)δ (...)*

Les diphones comportent maintenant un marqueur acoustique sur chacun de leurs phonèmes. Les paramètres acoustiques sont interprétés en temps réel en termes de variations par rapport aux valeurs de base, en utilisant les mêmes méthodes de modification par fenêtres sinusoïdales que lors de la normalisation. Ces valeurs sont à atteindre à chaque extrémité du diphone (correspondant à la partie stable de chaque phonème). Les valeurs intermédiaires correspondant aux autres périodes du diphone sont interpolées. Au fur et à mesure de sa construction, le signal de parole est envoyé à la carte-son qui le diffuse à la fréquence d'échantillonnage de 22 050 Hz. Pendant ce temps, le programme de l'utilisateur reste disponible pour d'autres tâches grâce à l'utilisation de processus simultanés.

8. Conclusion

Nous nous sommes ici attachés à décrire l'architecture d'un système de synthèse vocale à partir du texte en français dédié aux malvoyants. Ce programme de recherche appliquée nous a conduit à privilégier les critères d'efficacité, de fiabilité, d'évolutivité et de maintenance. Néanmoins, la nécessité de répondre à un problème applicatif concret avec des outils particuliers, n'occulte pas certaines questions majeures en recherche fondamentale, bien au contraire. Les problèmes posés par la génération automatique de la prosodie sont révélateurs à cet égard. Nos travaux nous ont ainsi permis de préciser l'apport des contraintes textuelles dans la construction de la structure intonative et ont fait émerger une hiérarchie dans les paramètres mobilisés (durée et déclinaison) pour la production d'unités distinctes comme les phrases et les paragraphes. Si cette stratégie d'analyse a des retombées pratiques évidentes en ce qui concerne l'agrément d'écoute, la concaténation de patrons prosodiques de phrases étant à l'origine de l'inévitable monotonie de la parole synthétique de la plupart des systèmes aujourd'hui, poursuivre dans cette voie ouvre nécessairement des perspectives de collaboration avec la linguistique textuelle et discursive. Concernant ensuite le lien entre syntaxe et prosodie tel qu'il a été formulé ici : le choix d'une grammaire qui repose sur le calcul de mises en relations syntaxiques, c'est-à-dire sur la mise en évidence de processus dynamiques en nombre finis à l'origine de la production ou de l'interprétation des structures, semble plus appropriée pour préciser les mécanismes sous-jacents à la construction prosodique que la description statique et peu économique de ces structures y compris exprimée sous la forme condensée de grammaires formelles. Dans cette approche, en effet, la phrase est considérée comme le codage linéaire d'une représentation dépendantielle abstraite, qui doit obéir à deux contraintes cognitives fondamentales : (i) la contrainte de minimisation de l'effort mémoriel implique la minimisation des distances entre deux nœuds syntaxiquement contigus dans l'ordre linéaire, (ii) l'information contenue dans la représentation profonde ne doit pas être perdue au cours de la linéarisation. La dimension temporelle associée implicitement à ces deux contraintes ne passe pas toujours, comme on l'a vu, par des restrictions distributionnelles. Dans de tels contextes, c'est bien en définitive la structure prosodique (temporelle et tonale) qui permet de répondre à ces principes : la variation dans le marquage et la durée des pauses, ainsi que dans le calcul des proéminences accentuelles permet d'évaluer la distance entre deux nœuds linéairement adjacents et ainsi de récupérer l'information nécessaire à l'auditeur pour reconstruire l'arbre de dépendance associé au matériau linéaire produit. La compression syllabique des mots grammaticaux nous fournit un autre exemple d'alignement entre les contraintes temporelles de nature rythmique et le traitement syntaxique. Selon plusieurs travaux menés sur le rythme du français [PAS 90] [DEL 95] [ZEL 98], les constructions prosodiques répondent à un principe de progression, selon lequel, les syllabes s'allongent progressivement du début à la fin

d'un constituant prosodique. Or, faire subir aux mots grammaticaux une compression syllabique par rapport à une durée moyenne de référence revient implicitement à appliquer ce principe, puisque les mots grammaticaux se situent essentiellement à l'initiale de constituants prosodiques. Pour conclure sur ce point, loin d'être antagonistes, les contraintes syntaxiques et rythmiques résultent de compromis et d'échanges pour actualiser de manière optimale la structure sonore du message.

En ce qui concerne la transcription graphème-phonème, la prise en compte de la hiérarchie linguistique des règles de prononciation nous a conduits à adopter une structure arborescente dès l'écriture des règles, dont les retombées pratiques en termes d'efficacité et de performance sont indiscutables : le fichier de règles s'agrandit rapidement en restant lisible et le taux de bonne transcription par phonème atteint à ce jour 99,66 %, plaçant d'ores et déjà notre système parmi les plus performants. Cependant, nous nous heurtons à un problème qui n'a pas reçu jusqu'ici de réponse théorique satisfaisante, à savoir la transcription des noms propres et mots d'emprunt, qui mobilise chez le lecteur des connaissances linguistiques étendues et multilingues, difficiles à théoriser et à modéliser. La seule alternative actuelle est strictement pragmatique : grâce à la collaboration des déficients visuels qui utilisent Kali, les erreurs de transcription sont recensées et ordonnées selon leur fréquence, ce qui facilite grandement le développement de nouvelles règles.

Le dernier problème que nous souhaitons soulever concerne la qualité intrinsèque de la voix synthétique, qui entre pour une large part dans l'intelligibilité et l'agrément d'écoute. Notre stratégie visait essentiellement l'efficacité, c'est-à-dire ici l'intelligibilité. Une évaluation comparative de notre système avec les principaux systèmes francophones, en cours de réalisation, nous a déjà confirmé l'intelligibilité remarquable de Kali. En revanche, l'agrément d'écoute reste assez moyen. Une réflexion sur les causes possibles de ce phénomène nous a conduits à effectuer des expériences destinées à isoler chaque étape de construction de la parole synthétique, remettant ainsi en cause certains de nos principes. Il semble que le point faible de notre voix synthétique réside dans l'utilisation de phonèmes types comme points cibles aux extrémités des diphones, notamment en ce qui concerne les voyelles. Dans la parole naturelle, en effet, une voyelle n'est pas toujours entièrement articulée : souvent obtenue par contraste avec les phonèmes qui l'entourent, elle n'a pas le même timbre qu'une voyelle prononcée isolément (phénomène de coarticulation). Ici, chaque voyelle est entièrement réalisée car il est nécessaire de pouvoir raccorder un diphone donné avec tous ceux qui possèdent le même phonème en interface. Cette méthode a pour conséquence une élocution un peu saccadée, articulée à l'excès, moins fluide que la parole naturelle. En contrepartie, la parole obtenue est très intelligible, même à vitesse d'élocution élevée. Une prosodie assez marquée, notamment au niveau du rythme, atténue en partie le caractère saccadé de la voix. Mais cette stratégie a ses limites et il nous semble maintenant indispensable de développer une méthode de mélange progressif en temps réel des diphones adjacents, qui devrait, pour une perte d'intelligibilité que nous espérons négligeable, produire une voix moins saccadée et ainsi améliorer grandement l'agrément d'écoute.

Enfin, notre travail en synthèse de la parole se poursuit sur la langue anglaise. Il nous permettra d'élargir les observations et les concepts décrits dans cet article et de mettre à l'épreuve le principe fondamental de notre démarche dans le traitement des données symboliques : la mise en place et l'exploitation d'une architecture informatique commune au traitement de différentes langues, seules les données et les règles manipulées étant sujettes à variation.

Références

[ABN 92] ABNEY S. (1992) : « Prosodic structure, performance structure and phrase structure », *Proceedings, Speech and Natural Language Workshop*, Morgan Kaufmann Publishers, San Mateo, CA, pp. 425-428.

[ALE 96] D'ALESSANDRO C., GARNIER-RIZET M., BOULA DE MAREÜIL P. (1996) : « Synthèse de la parole à partir du texte », *Fondements et perspectives en traitement automatique de la parole*, AUPELF-UREF, Henri Méloni (éd.), pp. 81-96.

[ALL 77] ALLEN J.B., RABINER L.R. (1977) : « A unified approach to short-time Fourier analysis and synthesis », *Proceedings, IEEE*, 65 (11), pp. 1558-1564.

[AUB 91] AUBERGE V. (1991) : La synthèse de la parole : des règles aux lexiques, Thèse de Doctorat, Université de Grenoble.

[BEL 92] BELRHALI R., LIBERT L., AUBERGE V., BOË L.J. (1992) : « Des lexiques aux règles : vers une méthode descriptive de la phonétisation du Français », *Actes des XIX^{èmes} Journées d'Études sur la Parole*, Bruxelles, Belgique, pp. 225-230.

[BEA 94] BEAUGENDRE F. (1994) : Une étude perceptive de l'intonation du français, développement d'un modèle et application à la génération automatique de l'intonation pour un système de synthèse à partir du texte, Thèse de Doctorat, Université de Paris XI, Orsay.

[BOU 97] BOULA DE MAREÜIL P. (1997) : Étude linguistique appliquée à la synthèse de la parole à partir du texte, Thèse de Doctorat, Université de Paris XI, Orsay.

[CHA 89] CHARPENTIER F., MOULINES E. (1989) : « Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones », *Proceedings Eurospeech '89*, ENST, Paris, vol. 2, pp. 13-19.

[DEL 95] DELAIS-ROUSSARIE E. (1995) : Pour une approche probabiliste de la structure prosodique, étude de l'organisation prosodique et rythmique de la phrase française, Thèse de Doctorat, Université de Toulouse-Le-Mirail.

[DEL 84] DELL F. (1984) : « L'accentuation dans les phrases françaises », F. Dell *et al.* (éd.), pp. 65-122.

[DUT 97] DUTOIT T. (1997) : *An introduction to text-to-speech synthesis*, Dordrecht, Kluwer Academic Publishers.

[GIG 97] GIGUET E., VERGNE J. (1997) : « From part of speech tagging to memory-based deep syntactic analysis », *International Workshop on Parsing Technologies 1997, proceedings*, MIT, Cambridge, MA, USA, pp. 77-88.

[GRA 14] GRAMMONT M. (1914) : *Traité pratique de prononciation française*, Paris, Delagrave.

[GRI 84] GRIFFIN D.W., LIM J.S. (1984) : « Signal estimation from modified short-time Fourier transform », *IEEE Trans. ASSP*, 32 (2), pp. 236-243.

[HAM 89] HAMON C., MOULINES E., CHARPENTIER F. (1989) : « A diphone synthesis system based on time-domain prosodic modifications of speech », *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, T. Durani éd., Glasgow, pp. 238-241.

[HIR 99] HIRST D., DI CRISTO A., ESPESSER R. (1999) : « Levels of representation and levels of analysis for the description of intonation systems », *Theory and experiment*, Dordrecht, Kluwer Academic Press, pp. 51-87.

[LAC 99] LACHERET-DUJOUR A., BEAUGENDRE F. (1999) : *La prosodie du français*, Paris, éditions du CNRS.

[LÉO 93] LEON P. (1993) : *Les phénomènes syntactiques : liaisons et enchaînements*, Paris, Nathan.

[MOR 81] MOREL M. (1981) : « Synthèse vocale par raccordement de segments d'oscillogrammes », *Revue d'acoustique du GALF*, n° 56, pp. 24-27.

[MOR 98] MOREL M., LACHERET-DUJOUR A. (1998) : « Utilisation d'une structure arborescente pour une hiérarchisation fine des règles de transcription graphème-phonème », *Actes des XXII*^{èmes} *Journées d'Etudes sur la Parole*, Martigny, Suisse, pp. 151-154.

[PAS 90] PASDELOUP V. (1990) : Modèle de règles rythmiques du français appliqué à la synthèse de la parole, Thèse de Doctorat, Université de Provence.

[ROS 99] ROSSI M. (1999) : *L'intonation, le système du français,* Paris, Ophrys.

[TES 59] TESNIERE L. (1959) : *Éléments de syntaxe structurale*, Paris, Klincksieck.

[VAN 99] VANNIER G. (1999) : Étude 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.

[ZEL 98] ZELLNER B. (1998) : Caractérisation et prédiction du débit de parole en français. Une étude de cas, Thèse de Doctorat, Université de Lausanne.