

L'intonation du discours : une implémentation par balises ; motifs et premiers résultats

Piet Mertens*, Antoine Auchlin[†], Jean-Philippe Goldman[†] & Anne Grobet[†]

* Département de Linguistique, K.U.Leuven

[†]Département de Linguistique, Université de Genève

ABSTRACT

Commonly assumed limitations of prosody generation in current text-to-speech systems based on the analysis of punctuation and syntax, are due to the fact that discourse and information structure is ignored. A strategy for obtaining more natural pitch contours is proposed, which is based on the insertion of tags indicating phonetic, acoustic, tonal, and/or functional attributes in the input text. A set of prosodic tags for speech synthesis is described.

1. INTRODUCTION

L'objectif général de notre travail est de parvenir à produire un signal de parole synthétique aussi naturel que possible en situation de TTS. A l'heure actuelle, les systèmes de TTS sont capables de *prononcer* des phrases (ou d'autres unités syntaxiques) ; mais dans un texte, les unités sont "en emploi", et leur réalisation vocale doit être aménagée afin de refléter prosodiquement "ce à quoi" chaque unité est *employée*, dans son environnement textuel : le système doit non seulement prononcer des unités linguistiques, mais faire comme s'il les employait. Or si les analyseurs peuvent déterminer la structure syntaxique d'une séquence de mots, ils ne peuvent pas reconstruire à partir de là les intentions et attitudes communicatives associées aux segments d'un texte. La pose de balises dans le texte vise fondamentalement à "forcer" l'intonation, afin qu'elle reflète l'intention communicative associée à l'emploi de chaque unité.

En tant qu'aspect de la communication parlée, l'intonation peut s'étudier sous plusieurs angles: celui de sa substance sonore, celui de sa forme, et celui de ses fonctions. Au niveau de la substance, l'intonation se manifeste par plusieurs propriétés sonores (changements de hauteur, accentuation, durée, débit, pauses, rythme, prises de souffle, qualité vocale). De plus, chaque langue se sert de son propre inventaire de formes intonatives et soumet leur utilisation à des contraintes syntaxiques. Mais il ne faut pas oublier que l'intonation a d'abord une fonction communicative: c'est un moyen linguistique pour transmettre certaines informations.

Quand le locuteur énonce un message, cette activité déclenche l'utilisation de certaines formes intonatives ainsi que l'ajustement des caractéristiques générales dont il était question plus haut. La réalisation des formes intonatives suppose à son tour des changements de hauteur et de force phonatoire situés à des points précis de la chaîne syllabique. L'expression d'une fonction pragmatique de

nature prosodique suppose ainsi l'activation d'une ou plusieurs formes intonatives précises (réalisées) dans une configuration prosodique donnée.

Dans leur tentative de générer des contours intonatifs adéquats à partir de la forme textuelle, les systèmes de synthèse vocale disposent seulement des informations présentes dans le texte – la ponctuation surtout, moins souvent des annotations indiquant la mise en valeur (italiques, caractères gras ou soulignement) ou la structure du texte (titre, autre, découpage en alinéas, paragraphes et sections) – ou de celles qui peuvent en être déduites, à savoir la structure syntaxique essentiellement. (En principe on pourrait envisager également le calcul automatique d'une représentation sémantique ; un tel objectif reste cependant un projet à long terme, pour ne pas dire futuriste.)

Or ces informations disponibles ne suffisent pas à obtenir des contours mélodiques expressifs, variés et naturels. Si l'on veut y arriver, il sera indispensable d'ajouter au texte des marqueurs pour signaler des aspects pragmatiques ou déclencher les formes prosodiques souhaitées. Ces marqueurs seront appelés des balises, par analogie avec les balises utilisées dans les documents hypertexte.

Dans la définition de ces balises, nous pouvons mettre à profit les observations faites plus haut sur la caractérisation de l'intonation à plusieurs niveaux d'analyse. Il est en effet utile de définir des balises pour chacun des niveaux de représentation : on aura ainsi des balises explicitant des aspects de substance (de nature acoustique), des formes (de nature symbolique) ou des fonctions. Nous reviendrons plus tard sur l'intérêt de cette approche.

Dans un premier temps on définit les balises "de bas niveau" et on vérifie leur bon fonctionnement : par exemple l'insertion d'une pause (silence) de durée spécifiée, la caractérisation de la tessiture de la voix (de synthèse) à partir de quelques paramètres. Ensuite on prévoit un deuxième ensemble de balises qui permettent d'imposer telle ou telle forme intonative à un endroit précis de la phrase : il peut s'agir du ton à utiliser en position accentuée finale, d'un accent d'insistance, etc. Enfin on arrive aux balises fonctionnelles, de nature linguistique, pragmatique, émotive, expressive ou (phono)stylistique ; pensons à la fin du tour de parole, à la mise en valeur d'un constituant, à l'expression de la consensualité, ou à un sentiment de colère, d'indifférence, d'angoisse, etc.

L'intérêt d'une telle approche réside dans son caractère modulaire et déductif. L'idée centrale est que la présence d'une balise fonctionnelle déclenchera l'insertion de balises de niveaux inférieurs, à savoir les marqueurs formels et acoustiques, selon une stratégie qui peut être formulée sous forme de règles. Une balise spécifiant un style emphatique, par exemple, entraînerait l'utilisation d'accents d'insistance et/ou du ton haut en position pénultième en fin d'énoncé. De la même façon les balises de nature émotive auront un effet sur les paramètres définissant la tessiture de la voix de synthèse, ou sur la présence des glissandos descendants en syllabe accentuée de fin de groupe intonatif. Comme la définition des balises fonctionnelles est indépendante des commandes formelles ou acoustiques et que les liens entre elles seront réglés à l'aide d'un ensemble de règles autonome, on obtient un système modulaire qui facilite l'expérimentation.

2. INFORMATIONS DISCURSIVES POUR LA POSE DES BALISES

2.1 Différentes dimensions discursives concernées

Au niveau discursif, la pose des balises fait suite à une pré-analyse du texte à intoner. Celle-ci doit fournir une interprétation extrayant certains des paramètres isolables selon l'approche modulaire de [Rou01]. Il ne s'agit pas de produire une analyse exhaustive (à supposer qu'on admette un tel concept), mais seulement d'extraire les aspects du discours susceptibles d'être spécifiquement reflétés intonativement. Dans le discours authentique spontané, l'intonation reflète aussi bien l'organisation informationnelle (thème/propos, premier/arrière-plan), la dimension hiérarchique et l'organisation relationnelle des actes de discours, l'organisation énonciative (voix représentées), que différentes « mimiques » à valeur affective et interactionnelle. Comme aux niveaux d'organisation inférieurs, les contraintes issues de ces différents niveaux peuvent être antagonistes et pas forcément satisfaites en même temps (une part du problème est de déterminer lesquelles, compte tenu du caractère partiellement opportuniste du marquage intonatif).

A un premier niveau configurationnel, la pré-analyse doit :

- identifier des liens privilégiés entre unités contiguës (regrouper), et des frontières correspondantes, à différents niveaux (hiérarchiser);
- identifier les éléments informationnellement et énonciativement focaux, le ou les "points", à différents niveaux de contraste;
- identifier les différentes instances énonciatives (locuteurs et énonciateurs) qui doivent être reflétées.

A ces éléments *configurationnels* s'ajoutent des facteurs *attitudinaux*, qui interfèrent les uns avec les autres et modulent de différentes façons l'organisation des données tirées des éléments configurationnels. Il s'agit pour

l'essentiel:

- de la force de l'engagement énonciatif associé au discours à tenir (non impliqué, "neutre", ou impliqué ;
- du degré d'adhésion ou de distanciation à l'égard du discours;
- de la tonalité affective (par prototypes Léo93 ou composants Sch89);
- du type d'attitude interactionnelle qui doit être adoptée (serviable et déférent, ou supérieur impérieux, détermine certains éléments importants de l'évolution de Fo (arrondi ou raideur des contours, etc. [Lad85]).

Nous avons traité prioritairement les données configurationnelles, pour diverses raisons techniques (facteurs articulatoires, précision et énergie, inaccessibles ; difficulté d'agir soûplement sur les variations de débit pour produire des regroupements rythmiques ; absence d'unités comme les prises de souffle dans les bases de diphtongues).

Les unités prosodiques qu'il s'agit d'entrer dans le texte correspondent grosso modo aux mouvements périodiques (m.p.) de [Rou01], unités intonatives présentées comme complètes et autonomes. Les m.p. sont constitués de un à n actes périodiques [Gro97]. ; ils peuvent s'appliquer à des unités syntaxiques de rang inférieur à la phrase (syntagmes majeurs) aussi bien qu'à des suites de phrases indépendantes ; s'ils sont bornés à gauche et à droite par les tours de parole, les mouvements périodiques peuvent s'intégrer en une projection prosodique maximale, équivalente à un échange. Notons que nous « chargeons » les unités *périodiques* de traits qui relèvent du statut *hiérarchique* des actes accomplis, comme le rapport de dépendance 'principal-subordonné', ou, au niveau supérieur, le trait 'initiatif-réactif' ; simplification sans conséquence ici.

2.2 Fonctions discursives

Si l'on suppose un m.p. formé de deux actes, l'un principal l'autre subordonné <AP-as>, la mise en relief prosodique de leur relation fonctionnelle peut être assujettie à la détermination de la « fonction illocutoire » spécifique de l'AP : si la paire <AP-as> est immédiatement constitutive d'un échange, la mise en relief de l'acte principal doit accentuer ou maximiser une marque spécifique de sa valeur d'emploi. Illustration :

Où étais-tu ? je t'ai cherché partout.

Par défaut, la traduction intonative du premier segment placerait un ton haut dynamique H/H (question) à la fin du premier segment « interrogatif » ; par défaut, si le balisage projetait les deux énoncés en un seul m.p., c'est le contour montant qui serait maximisé. Or cette suite peut parfaitement être comprise comme signifiant

[tu étais (caché) quelque part et je ne sais pas où ; je te le reproche parce que j'ai dû te chercher partout]

Si, dans son environnement textuel, la suite doit construire un tel sens, si la réponse est « excuse-moi », alors le premier segment, acte principal d'une intervention à

fonction illocutoire initiative de reproche, doit porter un contour descendant BB-, et le second segment être aménagé en conséquence.

2.3 Une heuristique paradoxale

On voit là que l'intonation est une véritable heuristique paradoxale. Dans le phénomène pudiquement nommé « analyse par la synthèse », la bonne intonation est à la fois le but et le guide ; elle est donnée avant qu'on en connaisse la ou les causes, alors qu'on voudrait maîtriser, comprendre, les causes afin de déterminer l'intonation. Bref : elle est évidente, transparente, mais opaque. On peut reprocher aux outils de synthèse utilisés ici d'être, avec leurs valeurs par défaut, peu aptes à restituer des faits discursifs ; ils n'en disposent pas moins, dès qu'on les manipule, d'un pouvoir de sur-analyse des données discursives tel que, d'un point de vue discursif, un grand nombre de manipulations disponibles ne peuvent purement et simplement pas être opérationnalisées ou fonctionnalisées de façon sérieuse.

3. LES SYSTÈMES DE BALISAGE EN SYNTHÈSE DE LA PAROLE

Notre volonté de manipuler la voix synthétique depuis le texte nous a orienté vers un système d'annotation du texte, qui soit à la fois hiérarchique et linéaire, pour décrire la structure discursive, syntaxique et prosodique. On le voudrait aussi extensible, c'est-à-dire qu'on pourrait ajouter de nouveaux types d'annotation.

3.1 Le balisage XML

Notre choix s'est porté sur le système de balisage XML qui est souvent jugé comme un généralisation de HTML, mais constitue en fait une simplification de SGML. Il est possible d'annoter un document avec des balises originales, au lieu d'avoir un jeu de balises prédéfinies et non-extensible. L'atout principal de XML est de faciliter le traitement automatisé de documents et de données. L'idée est de pouvoir structurer les informations de telle manière qu'elles puissent être à la fois présentées correctement à des lecteurs et traitées par des applications qui exploiteront de manière automatisée les informations en question.

Chaque balise peut contenir des informations additionnelles que l'on nomme attributs. Elle peut être **vide**, auquel cas elle est autonome, ou **non-vide** et se dédouble en deux balises dites ouvrante et fermante et un contenu. Ce dernier peut être du texte, d'autres balises ou les deux à la fois. La liste des balises autorisées et leur hiérarchie possible sont extensibles, et définies exactement dans une grammaire qui s'appelle une DTD (*document type definition*).

3.2 Les systèmes de balisage de la parole

Des sociétés ou des consortiums développant des systèmes de synthèse de la parole à partir du texte ont déjà établi des standards de balisage pour l'annotation de texte en vue d'une synthèse plus riche (XML SAPI, SABLE, JSML, VOICEXML).

Les grammaires de balises existantes sont très reliées à la théorie sous-jacente des systèmes de synthèse ainsi qu'à leur implémentation. En effet, les balises ne sont ni plus ni moins des commandes spécifiques qui vont modifier le déroulement par défaut de l'algorithme qui va générer la parole à partir du texte. Ces commandes invoquées par les balises sont prioritaires sur les étapes de l'algorithme. Certaines étapes sont peu flexibles, d'autres sont beaucoup plus malléables et configurables. Les balises peuvent les déclencher, les annuler ou simplement modifier certains paramètres.

Pour mieux se représenter le fonctionnement interne d'un système de synthèse de la parole, rappelons les étapes principales de la génération de la parole. Les systèmes sont souvent organisés en une cascade de modules au travers desquels transite le texte à synthétiser. Un système classique de synthèse de la parole déclenche d'abord un module de **traitement linguistique**, puis un module **phonétique** va transformer chaque mot (selon sa nature et son contexte linguistique) en une séquence phonétique. La génération de la prosodie peut se voir comme la succession de deux tâches: 1. une étape **phonologique** où l'on détermine les groupes intonatifs et l'accentuation de chaque syllabe (la syllabe est choisie comme unité rythmique); 2. une étape **acoustique** dans laquelle des spécifications prosodiques sont attribuées à chaque syllabe selon son accent et la frontière prosodique qui la suit. La synthèse du signal de parole est assurée par le **codeur**.

3.3 Proposition de système de balisage

Notre proposition est liée à l'implémentation du système de synthèse avec une description plus détaillée pour le module prosodique. Notre système est constitué d'un analyseur syntaxique fournissant une structure arborescente des syntagmes composant la phrase à prononcer. Ces informations détaillées sont utilisées par le module de phonétisation et par le module de génération de la prosodie. Ce dernier implémente un modèle théorique basé sur la superposition de la déclinaison globale et d'une composante d'accentuation des syllabes. De la structure syntaxique est déduite une structure prosodique puis des groupes intonatifs auxquels est attribué un accent final (et éventuellement un accent initial) symbolisé par un ton statique ou dynamique parmi plusieurs niveaux.

Nous présentons ci-dessous un ensemble de balises conçues selon les principes décrits plus haut et visant la synthèse de contours intonatifs variés et expressifs. Ces balises ont été en partie implémentées dans le système Mingus [M&a01].

4. Balises fonctionnelles

En principe, tout modèle tonal vise à rendre compte de la totalité des contours possibles d'une langue. Comme, en outre, chaque ton remplit une fonction précise, on pourrait, pour varier les contours en fonction du contenu à exprimer, se limiter aux balises tonales décrites plus haut. On ajoutera cependant un nouvel ensemble de balises, fonctionnelles cette fois-ci. L'intérêt de ces balises fonctionnelles réside dans leur indépendance vis-à-vis de la théorie pour la représentation des contours intonatifs.

Il est clair que l'utilisation de balises fonctionnelles entraîne une étape supplémentaire de conversion entre elles et les formes tonales ou acoustiques d'autre part. Afin d'illustrer cet aspect, cette conversion sera indiquée pour chacune des balises présentées.

Balises liées à la structure informationnelle

focus - La balise *<focus>* entoure le fragment textuel à focaliser par des moyens intonatifs. Dans le modèle tonal adopté, la focalisation correspond à l'utilisation du ton HL (ou HL-) en syllabe accentuée finale. Son effet est identique à celui de la balise *<tone af=HL>*.

e - La balise *<e>* (pour "emphase") provoque la mise en valeur du mot délimité par un accent d'insistance (ou accent initial, AI). Elle entraîne l'utilisation du ton haut (*<tone ai=H>*) sur la syllabe initiale du mot et l'insertion d'une pause avant cette même syllabe.

topic - La présence, en tête de phrase, d'éléments disloqués, de certains adjoints ou adverbiaux entraîne sur le plan intonatif une borne ou frontière infranchissable après ces éléments et dès lors une frontière intonative majeure, signalée par exemple par les tons HH ou /HH. Le phénomène est décrit par M. Rossi comme une "topicalisation". Il sera indiqué par la balise *<topic>*.

tail - La balise *<tail>* force un appendice sur l'entité entourée. Il s'agit d'une suite de syllabes atones à contour plat au niveau infra-bas ou haut, suivant le point d'arrivée de la syllabe accentuée finale précédente. Ce phénomène correspond à ce que Rossi appelle la "thématisation externe".

c'est de cela qu'il s'agit *<tail>* en quelque sorte
</tail>

Balises liées aux actes illocutoires

assert - La combinaison du ton final infra-bas et de la pénultième haute (soit le contour "...h L-L-") produit un effet assertif ou péremptoire. Ce contour sera déclenché par la balise *<assert>*.

question - La balise *<question>* provoque une intonation interrogative, soit une montée finale jusqu'au niveau haut rehaussé au cours de la syllabe accentuée finale. Elle a le même effet que *<tone af=H/H>*.

probe - La balise *<probe>* déclenche le ton haut en pénultième qui traduit un *léger appel de consensus*.

invite - La balise *<invite>* entraîne une montée tardive (LH) sous l'accent final qui traduit "l'interrogation avec appel de confirmation" (cf. Rossi).

smile - Cette marque indique l'effet de connivence que provoque le cliché mélodique correspondant à la séquence tonale "...h \HH", soit une pénultième haute suivie d'une syllabe accentuée finale à contour plat au niveau haut abaissé. Ce contour est parfois désigné sous le nom de "call contour" dans les travaux anglais.

Balises liées à la structure hiérarchique

parenthesis - Cette balise sera appliquée aux incises et parenthèses. Sur le plan prosodique elle se caractérise ordinairement par un passage local au registre bas. M. Rossi parlerait ici d'une "thématisation interne". Dans la parole expressive une variante au niveau haut se rencontre également. Les attributs low et high permettent de sélectionner l'une ou l'autre forme.

*je pense que c'est à ces problèmes nous devons
<parenthesis low> si naturellement monsieur
Fabius en est d'accord </parenthesis> consacrer
le débat de ce soir*

Balises liées à l'enchaînement des énoncés

link - La balise *<link>* indique un lien intonatif avec ce qui précède. Ce lien se manifeste par une attaque haute.

group - (*<MP>...</MP>*) groupe plusieurs phrases ou actes en un *mouvement périodique* dans lequel un contraste doit apparaître entre un acte principal et un argument. Le premier est précisé par la balise *<FRONT>* et le second par *<BACK>*. L'ordre de ces deux actes peut varier et l'effet principal de cette balise est le regroupement de plusieurs phrases sous le même gabarit de déclinaison.

Et encore...

Dans l'état présent des recherches, la liste des balises est loin d'être exhaustive. Il convient encore d'évoquer une balise existante liée à la langue choisie :

Language - indique un changement de langue. Comme mentionné plus haut, ce genre de balise peut avoir des répercussions sur plusieurs niveaux étant donné la spécificité des langues tant du point de vue phonétique (prononciation des mots), phonologique (syllable-timed vs. stressed-timed languages) ou même acoustique (certaines langues se distinguent par la largeur du registre mélodique).

Il reste enfin à élaborer des balises émotives et phonostylistiques qui contribuent elles aussi à donner une parole plus naturelle.

CONCLUSION

L'implémentation par balises de l'intonation du discours en est encore à ses débuts, mais elle montre bien quelles sont les voies à suivre : du côté de la synthèse, de nouveaux éléments (prise de souffle, par exemple) doivent

encore être implémentés pour rendre l'intonation plus naturelle. Du côté de l'analyse du discours, le recours à la synthèse constitue une heuristique qui permettra d'affiner les observations existantes.

BIBLIOGRAPHIE

- [Gro97] Grobet A. (1997) La ponctuation prosodique dans les dimensions périodique et informationnelle du discours, Cahiers de linguistique française 19, 83-123.
- [Lad85] Ladd D., Silverman K., Tolkmitt F., Bergmann G. & Scherer K. «Evidence for the independant function of intonation contour type, voie quality, and F0 range in signalling speaker affect» Journal of the Acoustical Society of America 78. 435-444.
- [Léo93] Léon P.R. (1993) Précis de phonostylistique, Paris, Nathan.
- [M&a101] Mertens, P., Goldman, J.-P., Wehrli, É. et Gaudinat, A. (à paraître) La synthèse de l'intonation à partir de structures syntaxiques riches TAL, 42/1.
- [Ros99] Rossi M. (1999) L'intonation, le système du français: description et modélisation, Paris, Ophrys.
- [Rou01] Roulet E., Filliettaz L., Grobet A., avec Burger, M. (2001) Un modèle et un instrument d'analyse de l'organisation du discours, Berne, Lang.
- [Sch89] Scherer K. (1989) "Les émotions: fonctions et composantes", in Rimé B. & Scherer K. (éds) Les émotions, Neuchâtel, Delachaux & Niestlé, 97-133.