Intonation du discours et synthèse de la parole : premiers résultats d'une approche par balises[1]

Piet Mertens*, Antoine Auchlin†, Jean-Philippe Goldman†,
Anne Grobet† & Arnaud Gaudinat†
* Département de Linguistique, K.U.Leuven
†Département de Linguistique, Université de Genève
<Piet.Mertens@arts.kuleuven.ac.be>

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[2]. 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.

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 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,

---

[1] Les exemples signalés par le symbole ♫ sont accompagnés de documents audio, qui peuvent être consultés sur la page des *Cahiers de linguistique française,* accessible depuis la page du Département de linguistique de l'Université de Genève : Http://www.unige.ch/lettres/linge/, rubrique « Publications ».

[2] Text-to-Speech.

nécessaires pour obtenir des contours mélodiques expressifs, variés et naturels. Pour 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. 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.

2. L'intonation : substance, forme et fonctions

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 entraîne ainsi l'activation d'une ou plusieurs formes intonatives précises dans une constellation prosodique donnée.

Dans la conception des 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 des balises explicitant des aspects de substance (de nature acoustique), des formes (de nature symbolique) ou des fonctions.

Dans un premier temps on définit ainsi 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 une émotion de colère, d'indifférence, d'angoisse, etc.

L'intérêt d'une telle approche réside dans son caractère modulaire et déductif. L'idée clé 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.

3. Informations discursives pour la pose des balises

3.1. 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 Roulet & al. (2001). 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 ; voir Grobet ici même), la dimension hiérarchique et l'organisation relationnelle des actes de discours, l'organisation énonciative (voix représentées), que différentes « mimiques » à valeur affective et interactionnelle[3]. 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).

---

[3] Voir par exemple le « sourire vocal » - Aubergé & Lemaître (2000) ; plus généralement Fónagy (1983), Léon (1993).

Ainsi, à un premier niveau, 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 relations, illocutoires[4] et interactives, qu'entretiennent les différents constituants les uns avec les autres ;
- identifier les différentes instances énonciatives (locuteurs et énonciateurs) qui doivent être reflétées.

Différents facteurs *attitudinaux* interfèrent avec ces éléments, et modulent de différentes façons l'organisation de leurs manifestations intonatives. 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 (qui peut être abordée en termes de prototypes sémiotiques, Léon 1993, ou de processus composants, Scherer 1989 - cette dernière approche étant cependant davantage compatible avec le cadre modulaire d'analyse du discours);
- du type d'attitude interactionnelle qui doit être adoptée (serviable et déférent, ou supérieur impérieux, détermine certains éléments importants de l'évolution de Fo, comme arrondi ou raideur des contours, etc. Ladd & al. 1985).

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

Les unités prosodiques qu'il s'agit d'entrer dans le texte correspondent grosso modo aux mouvements périodiques (m.p.) de Roulet & al. (2001), unités intonatives présentées comme complètes et autonomes. Les m.p. sont constitués de un à n actes périodiques (Grobet 1997) ; 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 des frontières de tours de parole (i.e. en fonction des informations provenant du module interactionnel), les mouvements périodiques peuvent s'intégrer en une projection prosodique

[4] Roulet (1980) considère l'intonation comme un « marqueur d'orientation illocutoire ».

maximale équivalente à un échange[5]. 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.

3.2. Fonctions discursives

3.2.1 Regrouper et contraster des « phrases »

L'exemple ci-dessous, vieil exemple de M. Martins-Baltar, permet d'illustrer un premier aspect:

(1) Il n'a pas plu. Le linge est sec

L'intonation de (1) doit d'une part présenter les deux énoncés comme entretenant une certaine relation, et d'autre part permettre de comprendre quelle est cette relation ou du moins orienter cette compréhension. Cela peut se faire notamment : i. en affaiblissant relativement la frontière qui sépare les deux phrases, et ii. en établissant un « contraste de contrastes », maximisant les marques prosodiques (intervalles, etc) sur le segment à mettre en valeur, et les minimisant sur l'autre. La zone syntaxique qui reçoit les contrastes forts peut être nommée « le point », et « contrepoint » la zone syntaxique qui lui est corrélée par affaiblissement des contrastes. Ainsi si le point est de parler du temps qu'il a fait, l'intonation placera notamment un accent majeur sur « pas plu » ; si le point est l'état du linge, on placera un accent sur « sec », et on affaiblira relativement les contrastes tonals de l'autre énoncé, le contrepoint. Point et contrepoint sont des « faisceaux » de propriétés, tant hiérarchiques, qu'informationnelles - topicales (Grobet ici même), « relationnelles » (relatives aux fonctions des constituants) ou encore énonciatives. L'alternance des formes intonatives sur les deux unités de (1) modifie simultanément ces différents aspects de leur interprétation discursive.

---

[5] Cette hypothèse va à l'encontre de celle de Nespor pour qui le phénomène de « restructuration », par lequel deux énoncés (courts) peuvent se présenter sous la forme d'un seul énoncé phonologique, ne saurait franchir de frontière de tour de parole. Du point de vue de sa forme intonative, la suite « A midi / impossible » peut servir à accomplir aussi bien un énoncé unique que deux énoncés brefs « restructurés », en une construction qui peut être aussi bien monologique (une intervention) que dialogique (un échange), et aussi bien monologale (une seule personne parle) que dialogale (deux personnes parlent). Ici, la modularité discursive permet d'affiner notablement les considérants des approches syntaxiques strictes.

Il faut noter que chacune des deux versions ainsi distinguables se distingue également de la version « pré-discursive », qui se contenterait d'intoner simplement les deux phrases successivement, comme deux assertions indépendantes ; il faut noter également que ce supplément n'est pas inférable des phrases qui composent le texte. Pour autant, il fait partie intégrante de l'interprétation la plus ordinaire du discours. Personne ne « se tromperait » d'intonation pour (1) si on l'entendait, dans un dialogue, en réponse à « où en est la lessive ? » ou « pourrai-je mettre le pantalon que tu as lavé ? » ; inversement, en réaction à « je me demande quel temps il a fait ici ce matin », personne n'intonerait la suite en pointant « le linge est sec ».

3.2.2 Fonction ou orientation illocutoire

Si l'on suppose un m.p. formé de deux actes, l'un principal l'autre subordonné, l'expression prosodique de leur relation fonctionnelle dépend du marquage de la fonction illocutoire du m.p. sur l'acte principal : la mise en relief de l'acte principal doit accentuer ou maximiser une marque spécifique de sa valeur d'emploi. Illustration :

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

Par défaut, la traduction intonative du premier segment placerait un ton haut montant 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 sur le premier acte qui serait maximisé. Or cette suite peut parfaitement être énoncée pour communiquer :

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

Si, dans son environnement textuel, la suite doit construire un tel sens - si par exemple un interlocuteur dit « excuse-moi » ensuite, alors le premier segment, acte principal d'une intervention à fonction illocutoire initiative de reproche, doit porter un contour descendant et une pente raide (ton HB- par exemple), et le second segment être aménagé en conséquence (ton d'appendice, par exemple).

3.3. Une heuristique paradoxale

Dans le traitement d'exemples de ce type, la manipulation empirique des formes intonative agit comme une véritable heuristique paradoxale : la manipulation entraîne des modifications interprétatives qu'il faut essayer de « capter » pour pouvoir saisir, en retour, ce qui est propre à la forme intonative. 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.

D'un côté, les outils de synthèse utilisés ici sont, par défaut, peu aptes à restituer crédiblement des séquences discursives, raison pour laquelle il est pertinent d'ajouter sous forme de balises différentes informations interprétatives « contextuelles[6] » ; mais d'un autre côté, ces outils n'en disposent pas moins, dès qu'on les met en oeuvre, d'un pouvoir de sur-analyse des données discursives : ils donnent accès à un grand nombre de manipulations, qui ne peuvent purement et simplement pas être opérationalisées ou fonctionnalisées de façon sérieuse, dans la mesure où l'on ne sait ni avec quoi les associer, ni comment.

4. 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.

4.1. Le balisage

Notre choix s'est porté sur le système de balisage XML[7]. Celui-ci facilite le traitement de documents et de données textuelles. L'idée est de pouvoir à la fois structurer, exploiter et présenter les informations de manière automatisée. L'atout principal du balisage de type XML est son extensibilité : il est possible d'annoter un document avec des balises originales, au lieu d'avoir un jeu de balises prédéfinies et non-extensible.

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. Voici quelques exemples :

---

[6] Au sens où elles ne peuvent être inférées du seul matériau verbal présent dans l'empan du traitement.

[7] Il est souvent jugé comme une généralisation de HTML, mais constitue en fait une simplification de SGML.

<phrase>…</phrase>	*balises non vides ouvrante et fermante*
<pause/>	*balise vide*
<pause durée=«250»/>	*balise avec attribut et valeur associée*

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*).

4.2. 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.

4.3. Proposition de système de balisage

Le jeu de balises proposées permet d'interagir dans le déroulement de chaque module du système de synthèse à partir du texte. Mais dans le cadre de cette étude, la plupart des balises décrites portent sur le module de génération de l'intonation.

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 les systèmes Mingus (Mertens & al. à paraître) et Fipsvox (Goldman ici-même).

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 le type d'accent et le ton associé à la syllabe ainsi que la frontière prosodique éventuelle qui suit. La synthèse du signal de parole est assurée par le **codeur**.

Les systèmes de synthèse FipsVox et Mingus sont constitués 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 sont attribués des tons. Puis des contours mélodiques et des durées de phonèmes et de pauses sont calculés.

4.3.1 Balises phonétiques

phono - permet de substituer directement la phonétisation d'un ou plusieurs mots. L'attribut **ph** permet de spécifier la transcription selon la convention SAMPA.

(4) <phono ph=«Z»> je </phono> crois que c'est indispensable

origine - précise l'origine linguistique. Le système consulte alors le lexique de la langue d'origine ou à défaut, charge un jeu de règles de phonétisation supplémentaire.

elision - indique au système le taux de réalisation des élisions.

liaison - indique au système le taux de réalisation des liaisons facultatives.

Les deux premières balises ont leur équivalent dans SABLE (<PRON ipa='' origin=''>). Il existe aussi <SAYAS> qui indique à la machine que le mot encapsulé est spécial et, selon sa nature, nécessite un traitement spécial pour une phonétisation correcte.

4.3.2 Balises formelles acoustiques

voice (voix) - Cette balise sélectionne la voix, c'est-à-dire la base de diphones, utilisée pour la synthèse vocale. Le nom de la base de diphones est spécifiée par l'attribut **db**.

(5) <voice db=«fr1»/>

grid (gabarit) - La tessiture caractéristique de chaque voix se définit par la gamme de hauteur comprise entre deux valeurs extrêmes, minimale et maximale, appelées parfois le plancher et le plafond. En synthèse vocale ces paramètres seront adaptés à la voix originale, afin d'éviter les voix dites métalliques. Cependant, dans la communication parlée, le locuteur se sert essentiellement d'une plage de hauteur au centre de sa tessiture; c'est le registre normal. Les variations mélodiques peuvent donc être caractérisées par un gabarit à quatre lignes: le plancher, le plafond et deux lignes pour le registre moyen. Celui-ci est délimité par des lignes (reconstituées à partir des valeurs observées en analyse, ou utilisées en synthèse), qui correspondent aux niveaux de hauteur bas et haut dans une représentation tonale. Elles sont appelées les lignes de déclinaison basse et haute. La distance entre ces deux niveaux correspond à l'ampleur des changements mélodiques (majeurs). Dans la parole lue le niveau bas descend au fur et à mesure que décroît la pression sous-glottique pour être remis à zéro après chaque prise de souffle. Dans la parole spontanée, il arrive que le niveau bas présente localement une pente ascendante, phénomène qui a une fonction expressive ou affective.

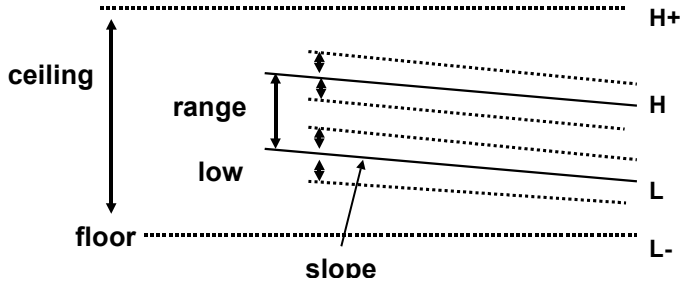

Figure 1. Définition paramétrique du gabarit mélodique.

Les paramètres sont 1. le plancher (floor): valeur minimale de la tessiture du locuteur (ou de la base de diphones correspondante), 2. l'intervalle mélodique entre ce plancher et le plafond (ceiling), 3. le niveau de hauteur bas (low) dans le registre habituel, 4. l'intervalle mélodique entre les niveaux bas et haut, soit l'ambitus (range), 5. la pente (slope) associée au niveau bas pour obtenir l'effet de déclinaison ou d'inclinaison, 6. l'intervalle mélodique mineur associé au rehaussement ou au rabaissement des niveaux de hauteur primaires.

Afin de contrôler tous ces aspects, la balise grid comporte plusieurs attributs. La forme <grid [range=R] [slope=S] [floor=F] [ceiling=C] [low=L]/> permet d'ajuster un ou plusieurs attributs du gabarit. La forme <grid reset> initialise les valeurs par défaut pour la voix activée.

register - Le changement de registre est caractéristique des incises et des parenthèses. Il se manifeste par un déplacement local, vers le bas ou vers le

haut, de la plage des variations mélodiques ainsi que par une diminution de leur ambitus. Le registre bas, de loin le plus fréquent, va généralement de pair avec une diminution du niveau sonore et avec une accélération du débit. Pour marquer correctement la frontière intonative associée au fragment affecté, sa dernière syllabe (accentuée) retourne déjà au registre initial. Le changement de registre est rendu par une modification locale du gabarit (grid). Ses effets sur le niveau sonore et le débit devraient également être pris en compte. (Mais le synthétiseur employé ne permet pas de moduler l'intensité, malheureusement.)

Le décalage du gabarit peut être exprimé de façon quantitative (décalage en demi-tons) ou catégorielle (choix entre les registres moyen, bas ou haut) :

(6) <register shift=«-4» range=4> .... </register>

(7) <register low> .... </register>

pause - La balise pause provoque un silence de la durée indiquée à l'endroit choisi. La durée peut être spécifiée de façon acoustique (en ms) ou catégorielle (choix entre une pause longue ou brève).

rate - De nombreux facteurs ont un effet sur le débit de parole : la durée de départ (ou moyenne) de chaque segment, son degré d'élasticité dans la syllabe, sa place par rapport à l'accent ou dans le groupe intonatif, l'effet allongeant des variations mélodiques, les pauses éventuelles, etc. Leur effet dépend de la sophistication du modèle de durée. Pour cette raison le choix d'un critère pour exprimer le débit est loin d'être évident. Il semble préférable de se limiter à un aspect, par exemple la durée segmentale par défaut, à laquelle on peut appliquer un coefficient d'allongement.

breathe - Jusqu'à récemment le rôle des prises de souffle a été entièrement négligé. Pourtant les respirations signalent souvent l'intention du locuteur d'entamer un tour de parole (Grobet & Auchlin ici-même). La plupart des synthétiseurs actuels ne prévoient pas de bruits de respiration. En attendant que ce manque soit comblé, on peut prévoir déjà la balise requise.

4.3.3 Balises formelles tonales

tone - Cette annotation permet de forcer sur le fragment textuel délimité un contour intonatif particulier, spécifié comme le choix d'un ton à la localisation indiquée. Les exemples suivants adoptent les tons et localisations proposés par Mertens.

(8) <tone af=HL> *démocratique* </tone>

(9) <tone ai=H> *démocratique* </tone>

(10) <tone penult=h> *démocratique* </tone>

(11) <tone ai=H penult=h af=«HL-»> *démocratique* </tone>

Le premier exemple provoque une chute du niveau haut au niveau bas sur la syllabe accentuée finale (AF) du mot. Le deuxième résulte en un accent initial (AI) haut sur la première syllabe du mot délimité. La troisième commande entraîne un ton haut à la syllabe pénultième (qui précède l'accent final dans le groupe en question). Enfin, le quatrième exemple spécifie les tons de plusieurs localisations à la fois.

Comme les balises se placent nécessairement aux frontières de mot, la place du ton dans le mot doit être déterminée par l'algorithme de génération de la prosodie. Les passages sans annotation tonale reçoivent les tons par défaut générés à partir des informations syntaxiques.

boundary - La balise <*boundary*> indique une borne dans le texte. Il s'agit d'un endroit qui sur le plan prosodique correspond à la frontière entre deux unités. Les éléments de part et d'autre de cette borne ne pourront pas être regroupés dans un même groupe accentuel ou intonatif. Cette balise facilite le traitement des incises ou des éléments disloqués, dans les cas où l'analyse syntaxique ne permet pas de les repérer.

4.3.4 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 à un certain niveau, 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 prosodique sous-jacente pour la représentation des contours intonatifs.

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

Il est clair également que cette étape de conversion fait apparaître à la fois différentes redondances (plusieurs formes intonatives étant susceptibles de remplir une même fonction), et différents amalgames fonctionnels (voir l’hypothèse du « faisceau » ci-dessus, une même forme accomplissant alors différentes fonctions simultanément). Dans une approche sémiotique ce serait une sérieuse limite. Dans un cadre modulaire, au contraire, ce sont là des effets de couplage entre différentes dimensions, qui restent certes à décrire, mais ne sont pas rédhibitoires, tant s’en faut.

4.3.4.1. 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>.

(12) je <focus> crois </focus> que c'est indispensable.

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.

(13) je pense que <e> démocratie </e> et débat politique

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 H/H. Le phénomène est décrit par Rossi (1999) comme une «topicalisation». Il sera indiqué par la balise <*topic*>.

(14) <topic> le vélo, le guidon </topic> il est cassé

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».

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

4.3.4.2. Balises liées aux enrichissements 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>. L'étiquette « question » est un terme générique pour une fonction illocutoire initiative manifestant qu'elle sollicite une complétion de la part de l'interlocuteur.

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).

allude - 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.

4.3.4.3. 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.

(16) je pense que c'est à ces problèmes nous devons consacrer <parenthesis low> si naturellement monsieur Fabius en est d'accord </parenthesis> le débat de ce soir (corpus Face-à-face Chirac - Fabius)

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 des actes peut varier ; l'effet principal de cette balise est le regroupement de plusieurs phrases sous le même gabarit de déclinaison.

4.3.4.4. Et encore…

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).

Dans l'état présent des recherches, la liste des balises est loin d'être exhaustive et définitive. Il reste à élaborer des balises émotives et phonostylistiques qui contribuent elles aussi à donner une parole plus naturelle. Il convient également d'approfondir l'étude des balises fonctionnelles, car il apparaît clairement qu'il n'existe pas de correspondance terme à terme entre fonction discursive et unité intonative, mais des liens entrecroisés entre, d'une part, une fonction discursive et plusieurs réalisations intonatives, et d'autre part, une même forme et plusieurs fonctions discursives.

5. Un exemple de balisage

Un tel article ne saurait se conclure sans une mise à l'épreuve des balises proposées ci-dessus. Dans ce but, nous avons utilisé un extrait de débat radiophonique que nous avons essayé de reproduire aussi fidèlement que possible en utilisant plusieurs procédés de synthèse vocale.

Dans un premier temps, l'extrait a été resynthétisé en préservant l'organisation temporelle (durée des sons et des pauses) et le tracé de hauteur d'origine. A cet effet un fichier de commandes de synthèse a été créé à partir d'une segmentation (effectuée par reconnaissance automatique de la parole) et de la copie des valeurs de F0 ; ces commandes sont ensuite envoyées au codeur MBROLA, qui est également utilisé par Fipsvox et Mingus. Ce premier type de synthèse fournit en quelque sorte la référence, l'objectif à atteindre en synthèse, puisqu'il montre les contraintes dues au codeur et les écarts dus à la différence de voix.

Dans un deuxième temps, on évalue le résultat en synthèse à partir du texte, où l'organisation temporelle et le contour mélodique sont générés par le traitement linguistique et prosodique, assistée par les balises ajoutées. Pour déterminer les balises nécessaires, nous avons soumis l'extrait à une première analyse perceptive, en suivant la méthode de notation proposée par Mertens et en nous aidant du logiciel Praat (pour la vérification des intervalles mélodiques). Cette première étape de l'analyse et ses résultats sont présentés de manière un plus détaillée dans l'article de Grobet (ici-même). A partir de cette première analyse, nous avons déterminé l'emplacement des balises formelles et fonctionnelles.

Le résultat se présente de la manière suivante : (la notation utilise les symboles suivants : # = pause, @ = prise de souffle, « ... « = diminution locale de l'intensité).

```
j' crois qu' c'est indispensable   j'entends par là j' pense que .   ###    @
     HL        b.........h /HH       b........b  HH  h...h    BB   (700ms)
   focus                                                            pause breathe

démocratie  et débat politique sont indispensables et je dirais     #    débat
 H h.\h BB « b.............b BB  b........../b H/H «  b......b B\B (150ms) H B\B
emph                                                                      emph

 ####   est indissociable de radio  et télévision de service public . ###
(860ms)  b  H  \h.b  B/B   b..h HH « b.......................b B-B- « (500ms)
pause       emph                                              fin

 @       je crois que le débat  #     à la radio  ou la télévision
              HH   \h..b. H HB (120ms) b..b. H.HH  «b..... b H/H «
breathe reset           emph
```

```
le débat politique en général mais le débat  à la radio et la télévision
  b..b/BB  H h..........\h HH  b....b  H\HH « b.....................b.H/H «
           emph                          emph

est l'occasion d' rapp'ler   ## @ que la politique ## @   c'est un débat d'idées
 b    H \h  BB \b..b   H/H (400ms)  b..b. H h  BH (300ms)   b....b  H.h....b BH
     emph                 pause breathe emph invite pause breathe emph    invite

c'est une confrontation de divergences une confrontation de points de vues
  b....b   H   h..b.BB   b....b  /B/B   b   H   h..b BB   b..............
           emph                            emph

différents
....b \BH
      invite
```

Nous avons réalisé deux exemples de synthèse de ce texte ainsi annoté, à l'aide de Fipsvox et de Mingus. (Le choix des balises varie selon les cas, selon les possibilités du système employé ; la forme exacte de l'entrée est donnée plus loin.) Il importe de bien souligner que le système de synthèse est responsable de l'ensemble des propriétés de la parole obtenue, jusqu'au moindre détail, en dehors de l'effet des balises rajoutées au texte.

La comparaison de la parole humaine et de ses imitations synthétiques met en lumière les points suivants.

1. La différence la plus saillante se situe au niveau du débit de parole, plus particulièrement le contraste entre le débit régulier de la parole de synthèse et la grande variation du débit dans la voix humaine. La voix naturelle donne l'impression d'un débit très élevé. Ceci est dû en partie aux nombreuses élisions de schwa dans les mots grammaticaux: «je crois que c'est indispensable...». Même quand on reproduit ces élisions en synthèse (grâce à la balise <phono>), la régularité du débit persiste pour la voix de synthèse. La parole humaine, en revanche, se caractérise par des accélérations et des ralentissements nombreux sur des fragments de parole de longueur variable, le plus souvent de l'ordre de plusieurs groupes intonatifs. La différence de débit n'est donc pas un problème lié au calcul de la durée des syllabes ou des sons individuels. Il suffirait en effet d'un coefficient lié au débit pour obtenir les effets souhaités. Seulement il n'est pas clair comment ce coefficient peut être obtenu, de quels facteurs il dépend.

2. Chez ce locuteur on est frappé par les variations d'amplitude au cours de la parole. Il ne s'agit pas ici de variations locales dues à l'accentuation de certaines syllabes, mais plutôt de diminutions de niveau sonore qui s'étendent sur un ou plusieurs groupes consécutifs et qui affectent l'ensemble des syllabes concernées, tant les syllabes atones que les accentuées. Dans l'extrait en question, les diminutions d'intensité semblent

remplir deux fonctions : indiquer (1) que le fragment affecté va avec ce qui précède, autrement dit, qu'il forme une unité plus grande avec le groupe de gauche; (2) qu'il s'agit d'une incise. Dans «démocratie <<et débat politique sont indispensables>> », le regroupement intonatif est déjà exprimé par les tons finaux ; on risque donc d'attribuer à l'intensité ce qui revient aux tons. Dans «de radio <<et télévision de service public>> » la diminution d'intensité n'est pas congruente à la structure des constituants. Le codeur MBROLA ne permettant pas de contrôler l'intensité, l'effet de la diminution d'intensité ne peut pas être reproduit en synthèse.

3. Le troisième écart récurrent qui se décèle à l'oreille concerne l'organisation temporelle à l'intérieur du groupe intonatif : la durée – des syllabes d'abord, mais également des certains sons – varie en fonction de l'accentuation, du type d'accent, du contour mélodique et de la place de la syllabe dans le groupe intonatif. Dans le modèle de durée mis en oeuvre dans Mingus et dans Fipsvox, le calcul de la durée syllabique prend en compte l'accentuation par l'accent final et la nature du ton (contour syllabique) à cette position. Les syllabes porteuses d'un accent initial, en revanche, sont traitées comme les syllabes atones, quant à la durée. Or, la comparaison des durées originales et calculées montre un effet évident de l'accent initial sur la durée de la consonne (ou du groupe consonantique) initial (onset). La présence d'un accent initial affecte également la durée des syllabes atones qui le suivent dans le groupe. Ces effets ne sont pas modélisés actuellement.

On ajoutera quelques remarques particulières, liées au système de synthèse employé.

L'entrée pour Mingus est donnée ci-dessous. Les balises utilisées se limitent à <tone> et <pause>.

(17) je <tone af=HL> crois </tone> que c'est <tone penult=h af=«/HH»> indispensable </tone> j'entends par <tone af=HH>là </tone> je pense que <pause len=«700»/> <tone ai=H penult=h af=LL> démocratie </tone> et débat <tone af=LL>politique</tone> sont <tone af=«H/H»>indispensables</tone> et je <tone af=«LL»>dirais</tone> <pause len=«150»/> <tone ai=H af=«LL»>débat</tone> <pause len=860> est <tone ai=H af=«L/L»>indissociable </tone> de <tone af=HH> radio </tone> et télévision de service <tone af=L-L->public </tone> ♬

Comme la balise <phono> pour la spécification de la transcription phonétique manque encore dans Mingus, la plupart des schwas sont prononcés. L'accentuation gênante de «j'entends» s'explique par une erreur d'analyse syntaxique : «j'entends» et «par là» sont considérés comme deux constituants juxtaposés sans relation syntaxique (et par conséquent deux

groupes intonatifs de même niveau), alors que la locution verbale «entendre par là» suppose leur intégration dans un même groupe ou au moins une subordination de l'un à l'autre. Cet exemple montre une fois de plus l'impact de la structure syntaxique sur l'intonation.

Le balisage de FipsVox s'appuie sur le logiciel FipsBalises (cf. infra § 7, et Goldman ici même) qui autorise une plus grande convivialité au niveau de la pose des balises.

(18) <PHONO ph=«Z»> je </PHONO> <PROSO af=«HL»> crois <PROSO break=«yes»> <PHONO ph=«k»> que </PHONO> c'est <PROSO af=«H»> indispensable j'entends par <PROSO af=«H»> là <PHONO ph=«Z»> je </PHONO> pense <PROSO af=«L»> que <PROSO pause=«700» break=«yes»/> <PROSO ai=«H»> démocratie </PROSO> et débat <PROSO af=«L»> politique </PROSO> sont <PROSO af=«HH+»> indispensables </PROSO> et je <PROSO af=«LL-»> dirais </PROSO> <PROSO pause=«150» break=«yes»/> <PROSO ai=«H» af=«LL-»>débat </PROSO> <PROSO pause=«860»/> <PHONO ph=«E»> est </PHONO> <PROSO ai=«H» af=«L» pause=«100»> indissociable </PROSO> de <PROSO af=«H»> radio </PROSO> <PROSO hauteur=«0.9»> et télévision <PHONO ph=«d»> de </PHONO> service public </PROSO> <PROSO pause=«500»/>

<PHONO ph=«Z»> <PROSO af=«H»> je </PROSO> </PHONO> crois que le <PROSO ai=«H» af=«HL»> débat </PROSO> <PROSO pause=«120»/> à la <PROSO ai=«H» af=«H»> radio </PROSO> ou la <PROSO af=«HH+»> télévision </PROSO> <PROSO debit=«0.8»> le débat politique en général mais le débat à la radio et la télévision </PROSO> est <PROSO ai=«H» af=«L»> l'occasion </PROSO> <PHONO ph=«d»> de </PHONO> <PHONO ph=«raple»> <PROSO af=«HH+»> rappeler </PROSO> </PHONO> <PROSO pause=«400»/> que la <PROSO ai=«H» af=«LH»>politique </PROSO> <PROSO pause=«300»/> , c'est un <PROSO ai=«H» af=«LH»> débat d'idées </PROSO> c'est une <PROSO ai=«H» af=«L»> confrontation </PROSO> <PHONO ph=«d»> de </PHONO> <PROSO af=«L»> divergences </PROSO> une <PROSO ai=«H» af=«L»> confrontation </PROSO> <PHONO ph=«d»> de </PHONO> points <PHONO ph=«d»> de </PHONO> vues <PROSO af=«LH»> différents </PROSO>. ♬

L'écoute de ces deux exemples de synthèse le montre clairement : chaque système a des points forts (par exemple élisions chez Fipsvox, marquage des deux grandes unités hiérarchiques, etc., bonne descente sur *public* avec Mingus, pour les deux, dans l'ensemble bon marquage des accents initiaux), mais aussi des points faibles - même limitées, toutes les balises ne peuvent pas encore être implémentées - qui représentent autant d'éléments à améliorer, en s'appuyant sur l'étude systématique de corpus plus étendus.

Quatre exemples sonores sont disponibles sur le site web indiqué dans le sommaire :

(19) **original** : extrait original ♬

(20) **prosocopie** : prosocopie de l'original, c'est-à-dire resynthèse par MBROLA à partir de l'alignement entre les mesures mélodiques et la segmentation automatique en phonèmes, après annotation des assimilations (« j'crois que »), élisions (« télévision d'service public ») et pauses. ♬

(21) **mingus** : synthèse par Mingus avec les balises issues de l'annotation prosodique ♬ (ex. 17)

(22) **fipsvox** : synthèse par FipsVox avec les balises issues de l'annotation prosodique ♬ (ex. 18).

6. 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 (intensité, 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.

7. Annexe : FipsBalises, un outil d'aide à l'annotation de texte pour la synthèse de la parole

Il s'agit d'un logiciel de synthèse de la parole se présentant sous la forme d'un traitement de texte rudimentaire dans lequel il est possible d'écrire un nouveau texte, d'en ouvrir un déjà existant, de le modifier et de l'enregistrer. Le synthétiseur FipsVox permet d'écouter le texte entier, une seule ligne ou bien une sélection.

Le principal attrait de cet outil est l'insertion aisée de balises durant la frappe, ou autour d'un morceau de texte sélectionné. Une fenêtre donne le choix parmi des balises prédéfinies et décrites ci-dessus, puis propose une liste d'attributs avec une valeur par défaut. L'utilisateur modifie les attributs qu'il désire puis valide ses choix pour insérer la balise dans le texte.

De plus, il est possible de créer ses propres balises à partir de une ou plusieurs balises pré-existantes. Ces macro-commandes ainsi définies simplifient et rendent plus lisible l'annotation en balises d'un passage.

Bien entendu, le synthétiseur tient compte des deux types de balises insérées (pré-définies ou définies par l'utilisateur) et modifie en conséquence la phonétisation et la génération de la prosodie.

Deux applications principales sont envisagées :

1. la synthèse d'un texte donné pour une diffusion grand public: une annotation *ad hoc* améliore substantiellement la synthèse par défaut qui ne reflètent pas toujours le ton voulu et comporte encore quelques défauts.

2. dans un but de recherche, cet outil permet de définir des balises fonctionnelles à partir des balises formelles. On vise ici le développement du jeu d'étiquettes possibles.

L'illustration ci-dessous présente une copie d'écran du logiciel FipsBalises, où l'on peut voir l'exemple décrit plus haut, la sortie phonétique et prosodique (fenêtre inférieure), la liste des balises (liste de droite et menu contextuel). Les balises <PHONO> et <PROSO> reprennent les balises définies plus haut, la balise <FOCUS> vient d'être créée.

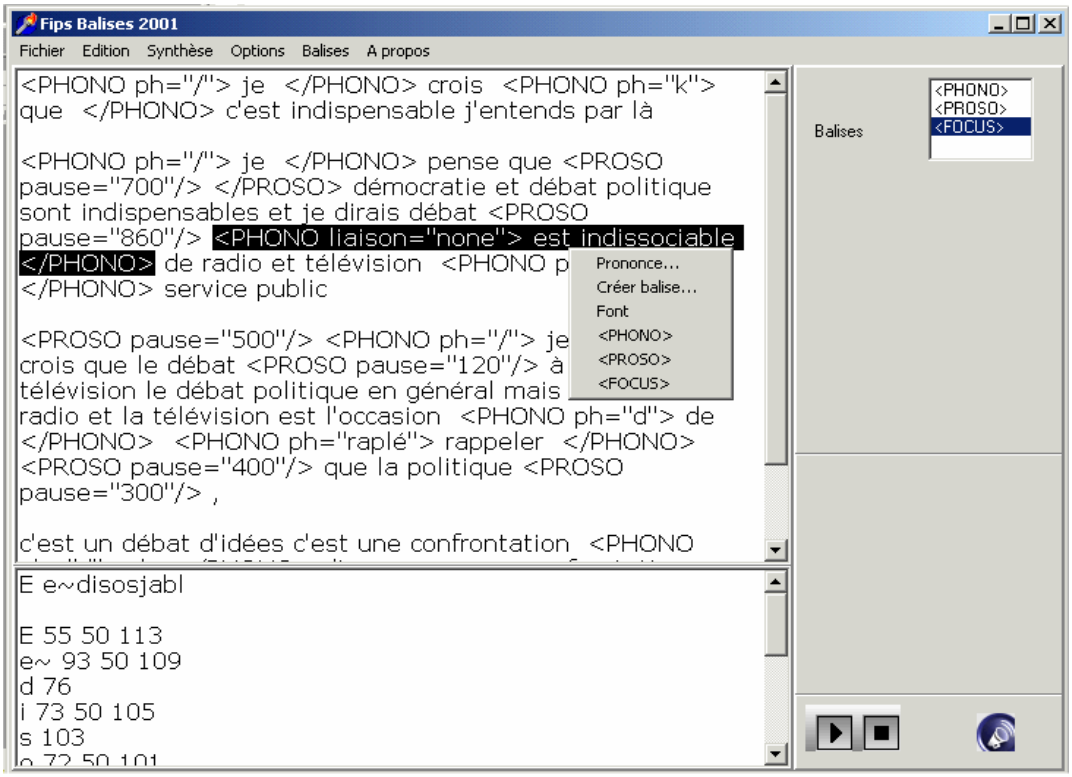

Figure 2 Capture d'écran de l'éditeur FipsBalises

Bibliographie

AUBERGE V. & L. LEMAÎTRE (2000), « The prosody of smile », ISCA Workshop on Speech and emotion, Newcastle, sept. 2000.

FONAGY I. (1983), *La vive voix. Essais de psycho-phonétique*, Paris, Payot.

GROBET A. (1997), « La ponctuation prosodique dans les dimensions périodique et informationnelle du discours », *Cahiers de linguistique française* 19, 83-123.

LADD D., SILVERMAN K., TOLKMITT F., BERGMANN G. & SCHERER K. (1985), « 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.

LEON P.R. (1993), *Précis de phonostylistique*, Paris, Nathan.

MERTENS P., GOLDMAN J.-P., WEHRLI E. & GAUDINAT A. (à paraître), « La synthèse de l'intonation à partir de structures syntaxiques riches », *TAL*, 42/1.

ROSSI M. (1999), *L'intonation, le système du français: description et modélisation*, Paris, Ophrys.

ROULET E. (1980), « Stratégies d'interaction, modes d'implicitation, et marqueurs illocutoires », *Cahiers de linguistique française* 1, 80-103.

ROULET E., FILLIETTAZ L., GROBET A., avec BURGER, M. (2001), *Un modèle et un instrument d'analyse de l'organisation du discours*, Berne, Lang.

SCHERER K. (1989), « Les émotions: fonctions et composantes », in RIMÉ B. & SCHERER K. (éds), *Les émotions*, Neuchâtel, Delachaux & Niestlé, 97-133.

SIMON A.C. & AUCHLIN A. (2001), « Multimodal, multifocal ? Les hors-phase de la prosodie », in Cavé C., Guaïtella I. & Santi S. (éds), *Oralité et gestualité. Interactions et comportements multimodaux dans la communication*, Paris, L'Harmattan, 629-633.